

# Introduction to Medical Statistics For Medical Students

---

Phichayut Phinyo

MD, Dip(ClinEpidemio) Dip(ClinStat)

Center for Clinical Epidemiology and Clinical Statistics

Department of Family Medicine

Faculty of Medicine, Chiang Mai University



# Class objectives

- After the end of this session, (hopefully) the students would be able to understand and correctly describe the concept of the following topics:

# Topics outline

- **Medical statistics I**
  - Statistical concepts in medical practice
  - Quantitative and Qualitative data
  - Types of variables
  - Frequency distributions
  - Measures of central tendency
  - Measures of dispersion

# Topics outline

- Medical statistics II
  - Statistical inference
  - Hypothesis generation and testing
  - The alpha level and p value
  - Type I error and type II error
  - Test of statistical significance
  - Test of clinical importance
  - Confidence intervals
  - Impact (size of effect)

# Topics outline

- Medical research methodology
  - Statistical tests
  - Internal and external validity
  - Statistical significance
  - Statistical power

# Nature of data

- Learning objectives
  - Variables and Data
  - Nominal data
  - Ordinal data
  - Discrete metric data
  - Continuous metric data
  - Scale of measurements



# Variables and data

- **Variables:** something whose value can vary
- **Data:** value that explains the variable

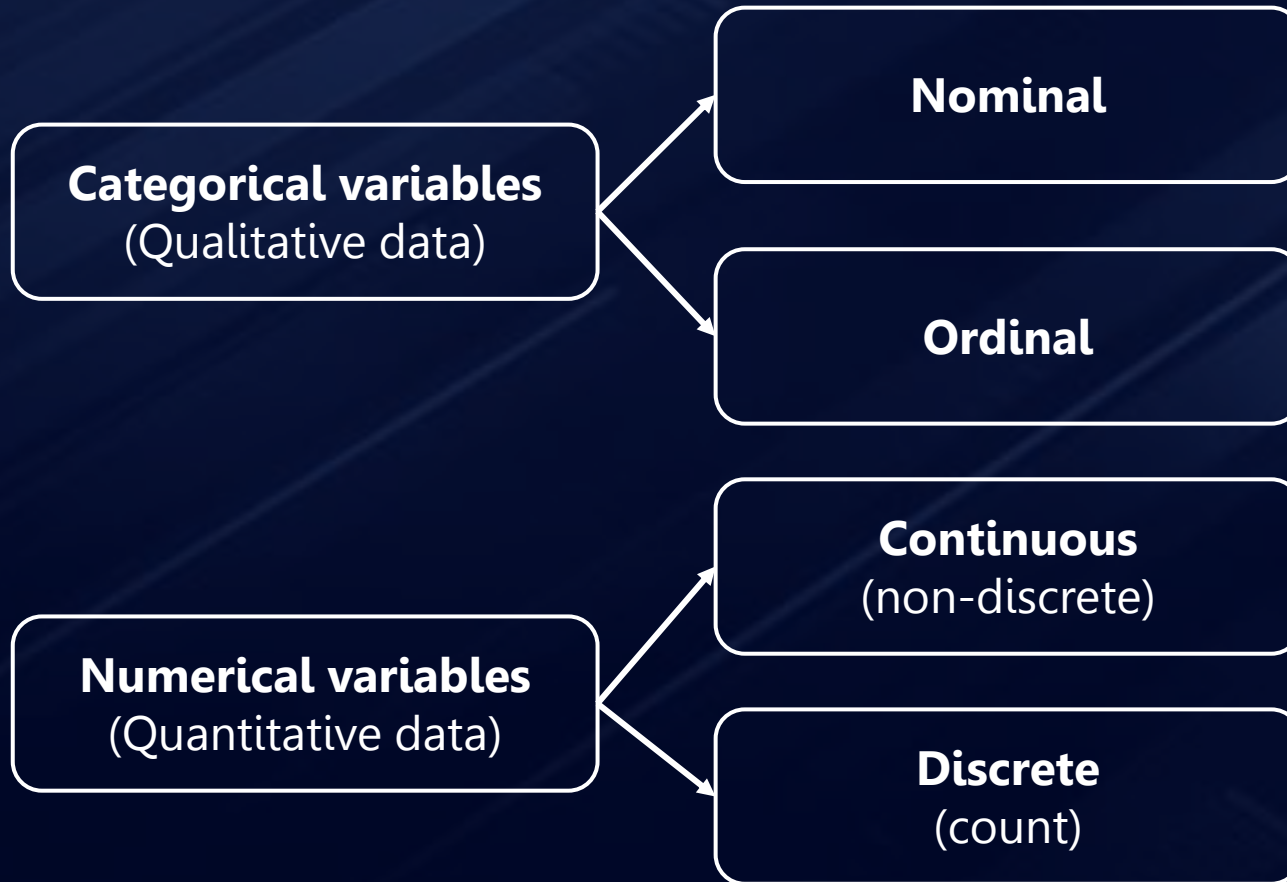


## Variables

- Age
- Gender
- Zodiac signs
- Glasses
- Etc.

We get the data when we determine the value of a variable for each unit of observation

# Types of variables (data)





# Types of variables (data)

- **Nominal data**

- Related to named things
- Particularly not numeric, but categories
- Can be more than 2 categories
- The ordering is arbitrary (or does not make sense)

- **Example**

- Gender (male or female)
- Ethnicity (Thai, Burmese, Laos, etc.)
- Zodiac signs (Cancer, Virgo, Sagittarius, etc.)
- Insurance (UC, SSSS, CSS, etc.)
- Religion (Buddhism, Islam, Christian, etc.)
- Marital status (Single, Married, Divorce, etc.)

# Types of variables (data)

- **Nominal data**

- Related to named things
- Particularly not numeric, but categories
- Can be more than 2 categories
- The ordering is arbitrary (or does not make sense)

- **Clinical example**

- Location of headache (diffuse, temporal, etc.)
- Mechanism of injury (traffic, falling, etc.)
- Site of metastasis (brain, bone, liver, etc.)
- ...

# Types of variables (data)

- **Ordinal data**
  - Usually are assessed measurements
  - Not numerical data but categorical
  - Can be more than 2 categories (usually  $\geq 2$ )
  - The ordering is not arbitrary (make sense to order)
- **Example**
  - Cancer staging (stage I, II, III, IV)
  - Disease severity (mild, moderate, severe)
  - Risk evaluation (low, intermediate, high)
  - Glasgow Coma Scale (GCS)
  - Motor power (grade 1,2,3,4,5)

# Types of variables (data)

- Ordinal data
  - Special consideration about ordinal data
  - The ordinal scales **are not real number** but only **numeric labels** that were attached to each category.
  - The difference between each value might not be proportional. In other words, we cannot quantify the **interval difference** of ordinal data.
  - **Not appropriate to apply any rules of basic arithmetic to this type of data.** You should not add, subtract, multiply, divide, or even find the average value of ordinal data.

# Types of variables (data)

- Discrete metric data (discrete data or count data)
  - Comes from counting process
  - Unit of measurements: *number of things*
  - Real numbers (integers)
  - Decimals do not make sense
  - Equal interval difference, unlike ordinal
- Clinical example
  - Number of students in the classroom
  - Number of patients
  - Number of deaths
  - Number of angina attacks
  - Number of hospital visits

# Types of variables (data)

- Continuous metric data (continuous data)
  - Comes from measurements
  - Have units of measurement
  - The data are real numbers
  - Decimals do have meaning
  - Equal interval difference
- Clinical example
  - Body weight, height, body mass index
  - Blood sugar, Serum cholesterol level
  - Hemoglobin level
  - Tumor marker level
  - Blood pressure (continuous or discrete?)
  - Heart rate (continuous or discrete?)



# Types of variables (data)

- What is the main difference between discrete and continuous metric data?



- (a) List the possible number of eggs that the carton contain  
(counting)
- (b) List the number of possible value for the weight of each egg in the carton  
(measuring)

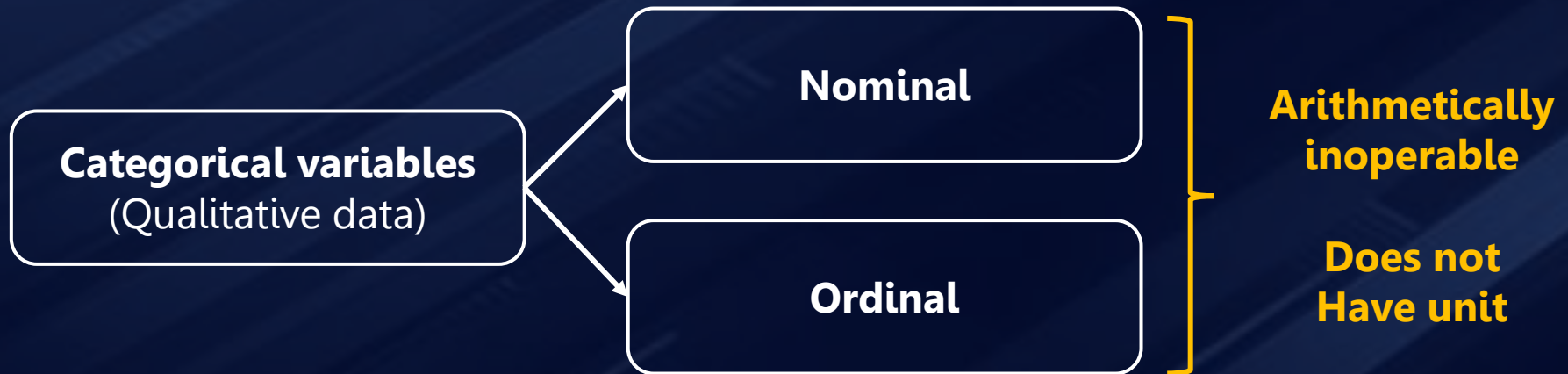
# Types of variables (data)

- What is the main difference between discrete and continuous metric data?

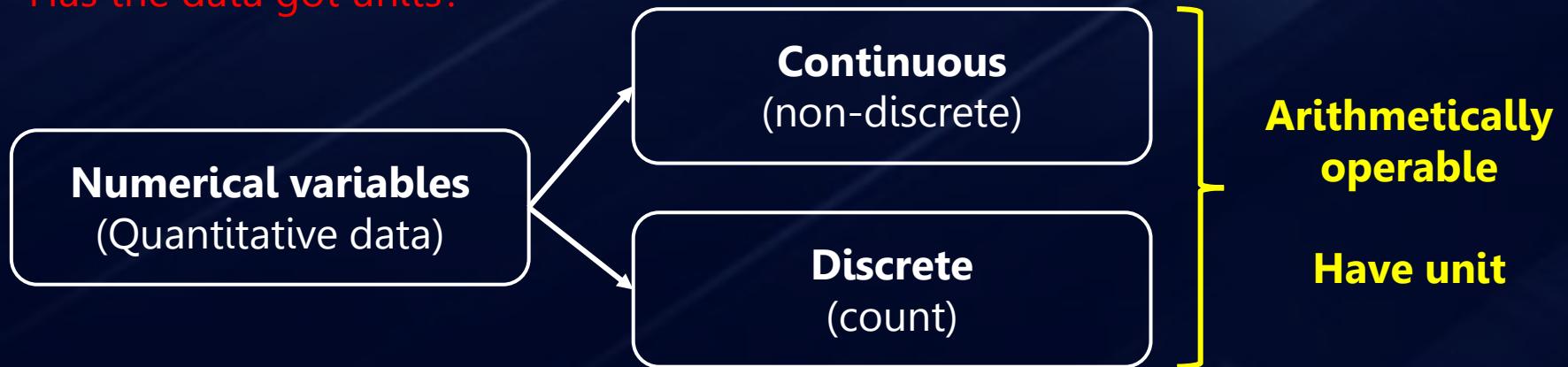


- (a) List the possible number of eggs that the carton contain  
(counting 0,1,2,3,4,5,6)  
limited range of possible data
- (b) List the number of possible value for the weight of each egg in the carton  
(measuring 70,70.1,70.001,70.000001, 70.0000000000000001 etc.)  
Infinite range of possible data

Can the data be put in meaningful order?



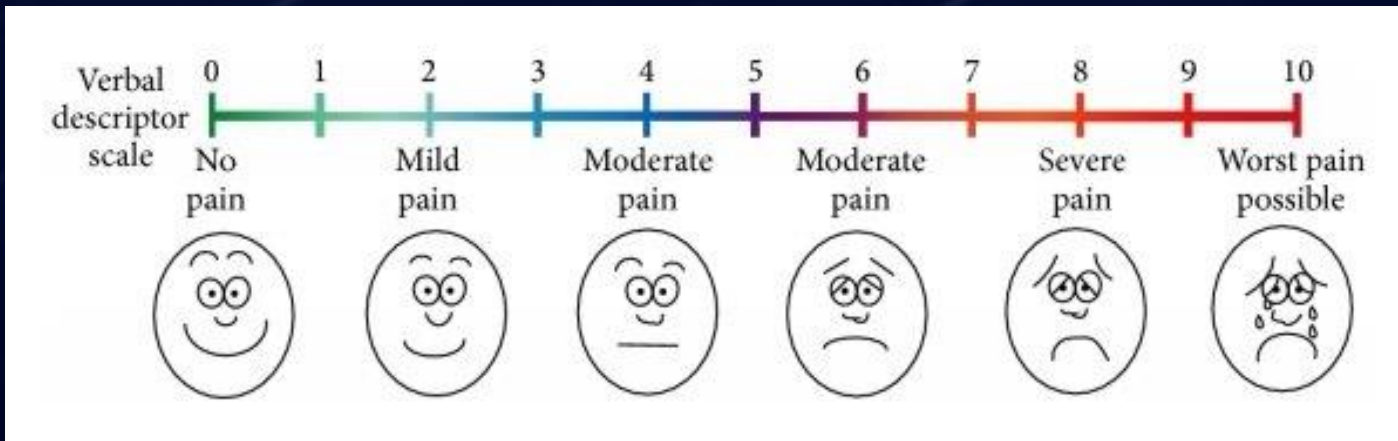
Has the data got units?



Do the data come from counting or measuring?

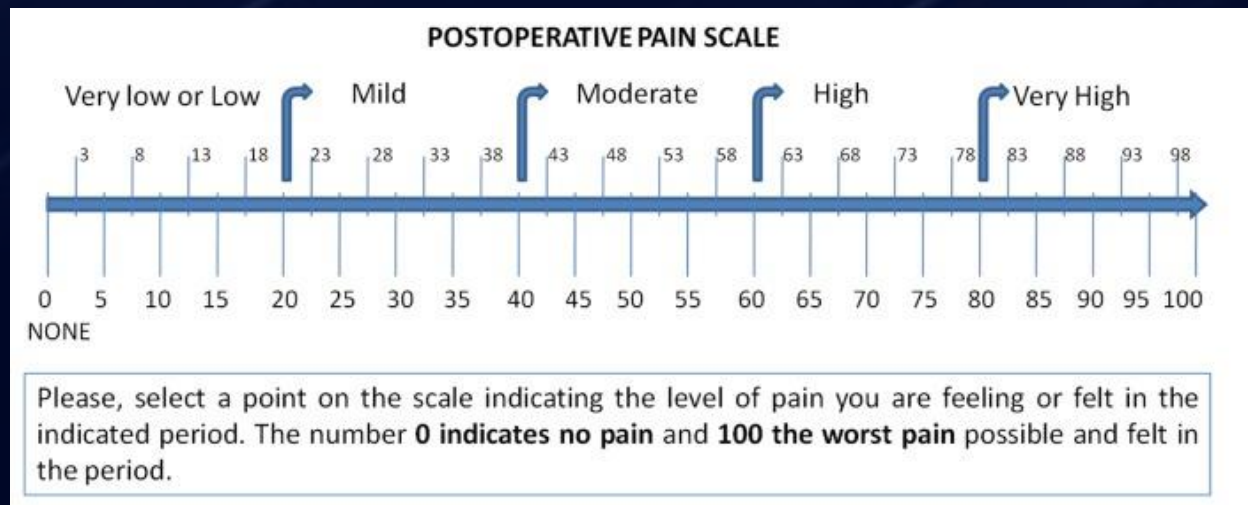
# Quizzes

- **Numeric rating scale (NRS)**
  - What types of data is it?
    - Nominal
    - Ordinal
    - Discrete metric
    - Continuous metric



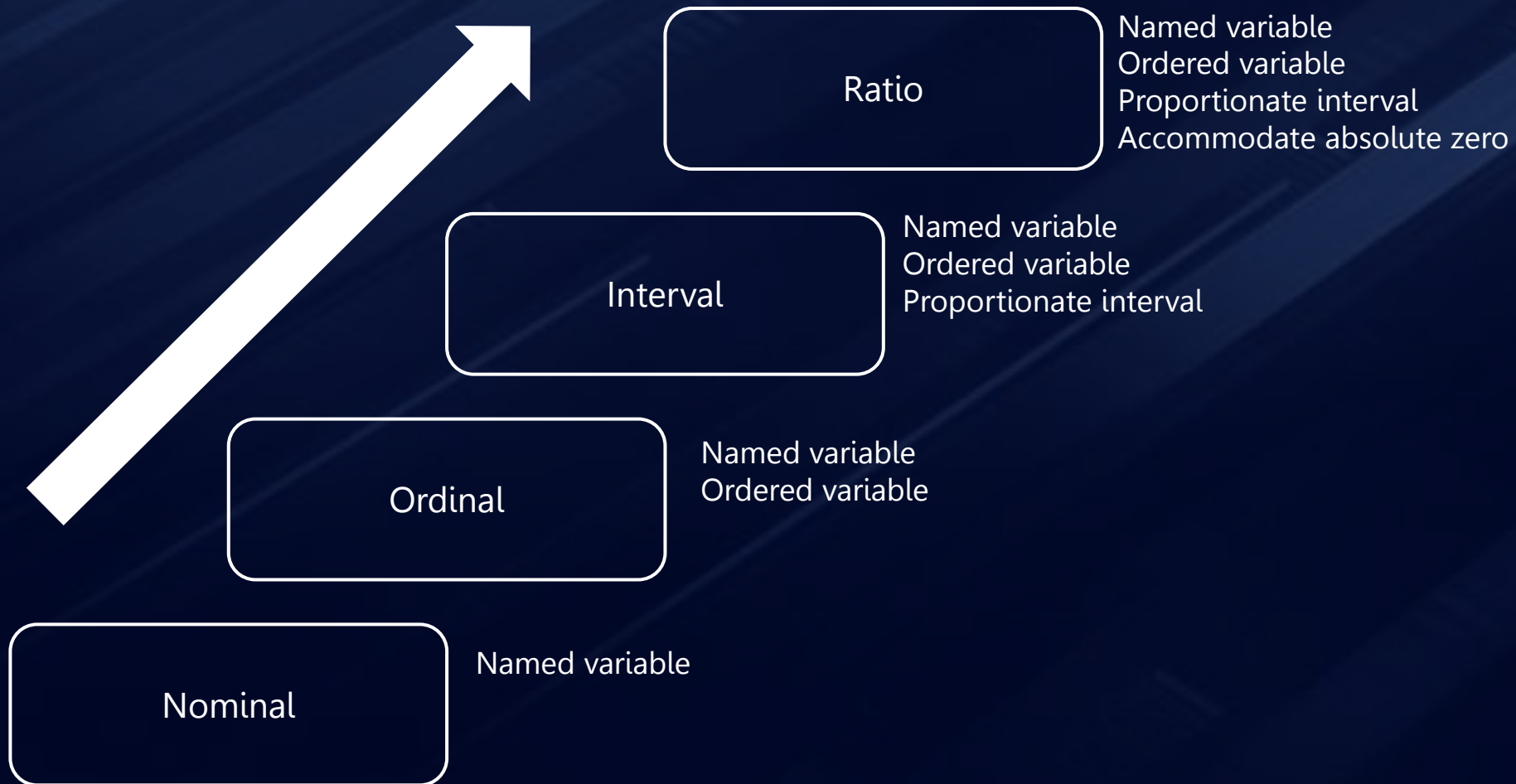
# Quizzes

- Visual Analogue Scale
  - What types of data is it?
    - Nominal
    - Ordinal
    - Discrete metric
    - Continuous metric





# Scale of measurement





# Describing data

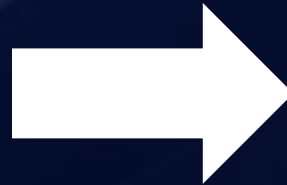
- Learning objectives
  - Frequency distribution
  - Frequency table
  - Relative frequency
  - Cumulative frequency
  - Relative cumulative frequency
  - Grouped frequency tables

# Raw data: GENDER (nominal)

M	F	F	F	M	F	F	M	M	F
F	F	M	M	F	F	F	M	F	F
F	M	F	F	M	M	M	M	M	F
M	M	M	F	M	M	M	M	F	F
M	F	M	M	F	F	M	M	F	M
F	F	F	M	M	M	M	F	F	F
F	F	M	M	F	F	F	M	M	M
M	M	F	F	M	M	M	M	M	F

# Raw data: GENDER (nominal)

M	F	F	F	M	F	F	M	M	F
F	F	M	M	F	F	F	M	F	F
F	M	F	F	M	M	M	M	M	F
M	M	M	F	M	M	M	M	F	F
M	F	M	M	F	F	M	M	F	M
F	F	F	M	M	M	M	F	F	F
F	F	M	M	F	F	F	M	M	M
M	M	F	F	M	M	M	M	M	F



Data Editor (Edit) - [Untitled]

File Edit View Data Tools

var8[11]

	id	gender
1	1	M
2	2	F
3	3	F
4	4	M
5	5	M
6	6	F
7	7	F
8	8	M

# Frequency tables (nominal)

Gender	Count (Frequency)	Percent	Cumulative percentage
Female	38	47.5	47.5
Male	42	52.5	100
Total	80	100	

To describe nominal data, **frequency (percentage)** tables are used. The order in the table is not important.

# Frequency distributions

<b>Insurance</b>	<b>Count (Frequency)</b>	<b>Percent</b>	<b>Cumulative percentage</b>
UC	50	50.0	50.0
SSSS	30	30.0	80.0
CSS	15	15.0	95.0
Others	5	5.0	100
Total	100	100	

Even though ordering of nominal categories is arbitrary, arranging the category from highest frequency to lowest frequency could help the readers understand the data more quickly.

# Frequency tables (nominal)

Number (frequency)

Insurance	Count (Frequency)	Percent	Cumulative percentage
UC	50	50.0	50.0
SSSS	30	30.0	80.0
CSS	15	15.0	95.0
Others	5	5.0	100
Total	100	100	

Even though ordering of nominal categories is arbitrary, arranging the category from highest frequency to lowest frequency could help the readers understand the data more quickly.



# Frequency distributions

Percentage (relative frequency)

Insurance	Count (Frequency)	Percent	Cumulative percentage
UC	50	50.0	50.0
SSSS	30	30.0	80.0
CSS	15	15.0	95.0
Others	5	5.0	100
Total	100	100	

Even though ordering of nominal categories is arbitrary, arranging the category from highest frequency to lowest frequency could help the readers understand the data more quickly.

# Frequency distributions

Does this make sense?

Insurance	Count (Frequency)	Percent	Cumulative percentage
UC	50	50.0	50.0
SSSS	30	30.0	80.0
CSS	15	15.0	95.0
Others	5	5.0	100
Total	100	100	

Even though ordering of nominal categories is arbitrary, arranging the category from highest frequency to lowest frequency could help the readers understand the data more quickly.

# Frequency tables (ordinal)

Cancer stage	Count (Frequency)	Percent	Cumulative percentage
Stage I	30	12.0	12.0
Stage II	50	20.0	32.0
Stage III	100	40.0	72.0
Stage IV	70	28.0	100.0
Total	250	100	

Percentage  
Cumulative  
Frequency  
Or  
Relative  
Cumulative  
Frequency

Ordering of categories is not arbitrary and is meaningful for ordinal data, such as this case. The frequency distribution shown here would reflect the distribution of cancer stages in the sample.

# Frequency tables (discrete)

Parity	Count (Frequency)	Percent	Cumulative percentage
0	49	49.0	49.0
1	18	18.0	67.0
2	17	17.0	84.0
3	11	11.0	95.0
4	2	2.0	97.0
5	1	1.0	98.0
≥6	2	2.0	100
Total	100		

# Frequency tables (continuous)

<b>Hemoglobin</b>	<b>Count (Frequency)</b>	<b>Percent</b>	<b>Cumulative percentage</b>
8.80	1	1.0	1.0
8.90	1	1.0	2.0
8.91	1	1.0	3.0
8.92	1	1.0	4.0
8.93	1	1.0	5.0
9.01	1	1.0	6.0
9.02	1	1.0	7.0
...	...	...	100
Total	100		

# Frequency tables (continuous)

Hemoglobin	Count (Frequency)	Percent	Cumulative percentage
8.80	1	1.0	1.0
8.90	1	1.0	2.0
8.91	1	1.0	3.0
8.92	1	1.0	4.0
8.93	1	1.0	5.0
9.01	1	1.0	6.0
9.02	1	1.0	7.0
...	...	...	100
Total	100		

**Need to be  
grouped**

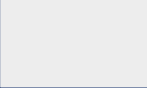

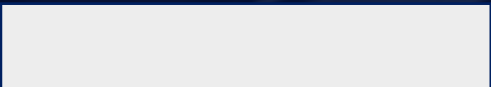

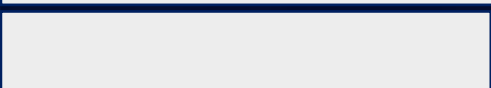




# Grouped frequency table

<b>Hemoglobin</b>	<b>Count (Frequency)</b>	<b>Percent</b>	<b>Cumulative percentage</b>
7.0-8.0	10	6.5	6.5
8.0-9.0	20	12.9	19.4
9.0-10.0	30	19.4	38.8
11.0-12.0	35	22.4	61.2
12.0-13.0	30	19.4	80.6
13.0-14.0	20	12.9	93.5
14.0-15.0	10	6.5	100
Total	155		

General rule of thumb: no lesser than five and no more than 10 groups

# Grouped frequency table

Hemoglobin	Count (Frequency)	Percent	
7.0-8.0	10	6.5	
8.0-9.0	20	12.9	
9.0-10.0	30	19.4	
11.0-12.0	35	22.4	
12.0-13.0	30	19.4	
13.0-14.0	20	12.9	
14.0-15.0	10	6.5	
Total	155		

# Grouped frequency table

Hemoglobin	Count (Frequency)	Percent
7.0-8.0	10	6.5
8.0-9.0	20	12.9
9.0-10.0	30	19.4
11.0-12.0	35	22.4
12.0-13.0	30	19.4
13.0-14.0	20	12.9
14.0-15.0	10	6.5
Total	155	

Around 60% are  
among these categories  
(from 9.0-13.0)



# Frequency distribution

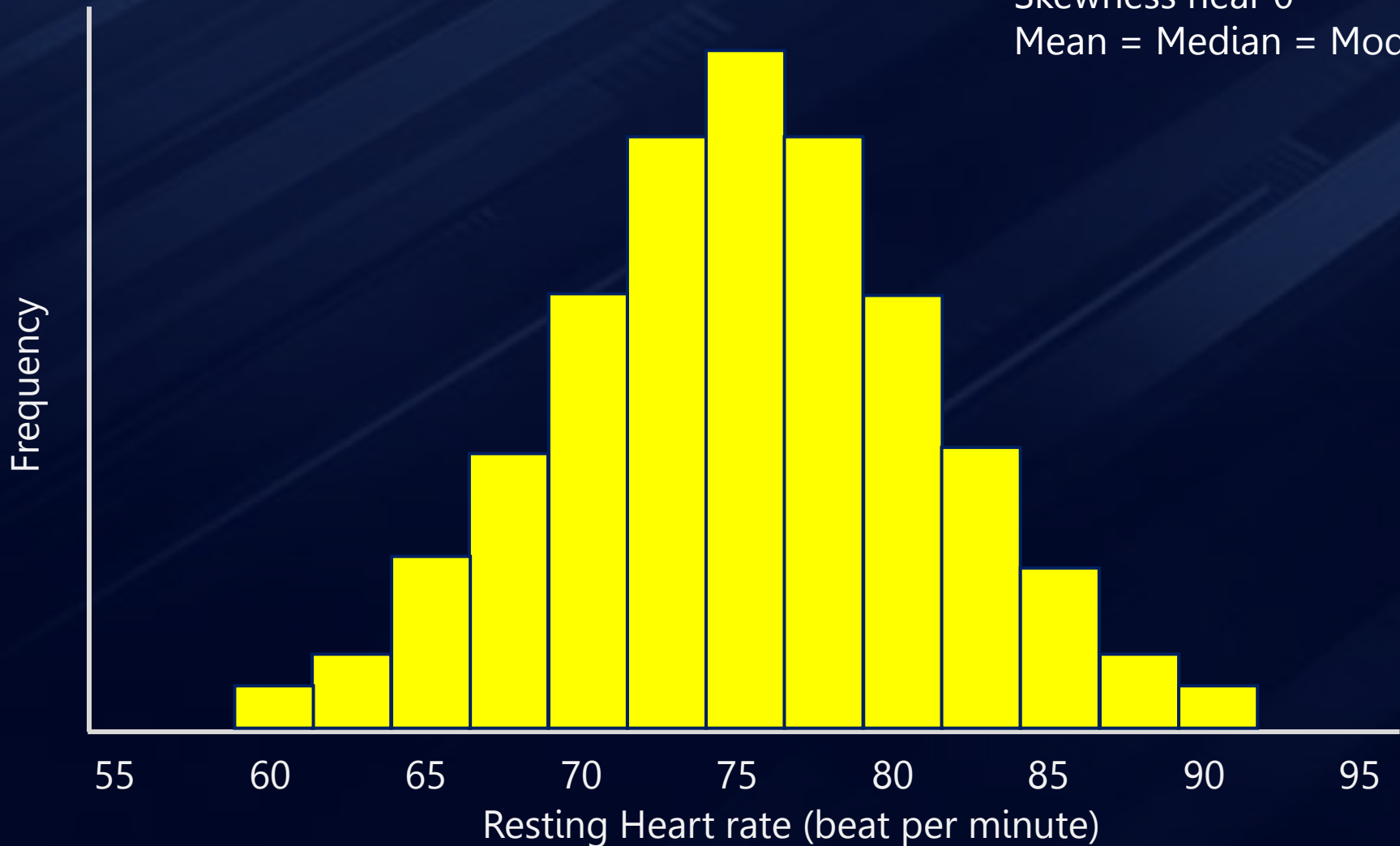
- Commonly illustrated via histogram
- Measure of shape
  - Skewness
  - Kurtosis
- Common distribution
  - Normal distribution
  - Skewed distribution
  - Uniform distribution
  - Bimodal or multimodal distribution

# Frequency distribution

- Measure of shape
  - Skewness
    - Measure of the symmetry
    - Measured by skewness coefficient
    - Range from -1 and +1
    - When the distribution is symmetry, the skewness coefficient will be near 0

# Histogram

Symmetrical distribution  
Normal distribution  
Skewness near 0  
Mean = Median = Mode

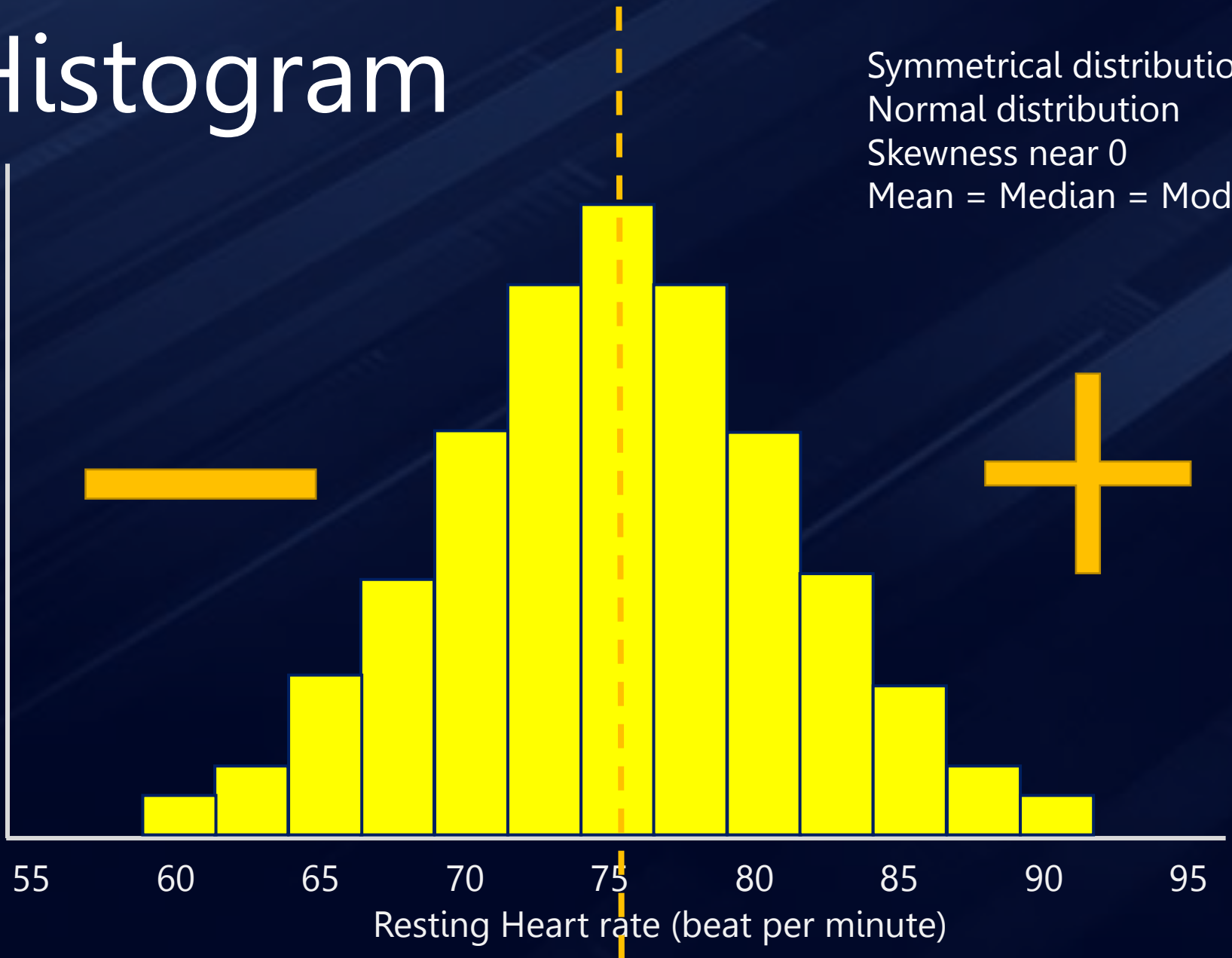




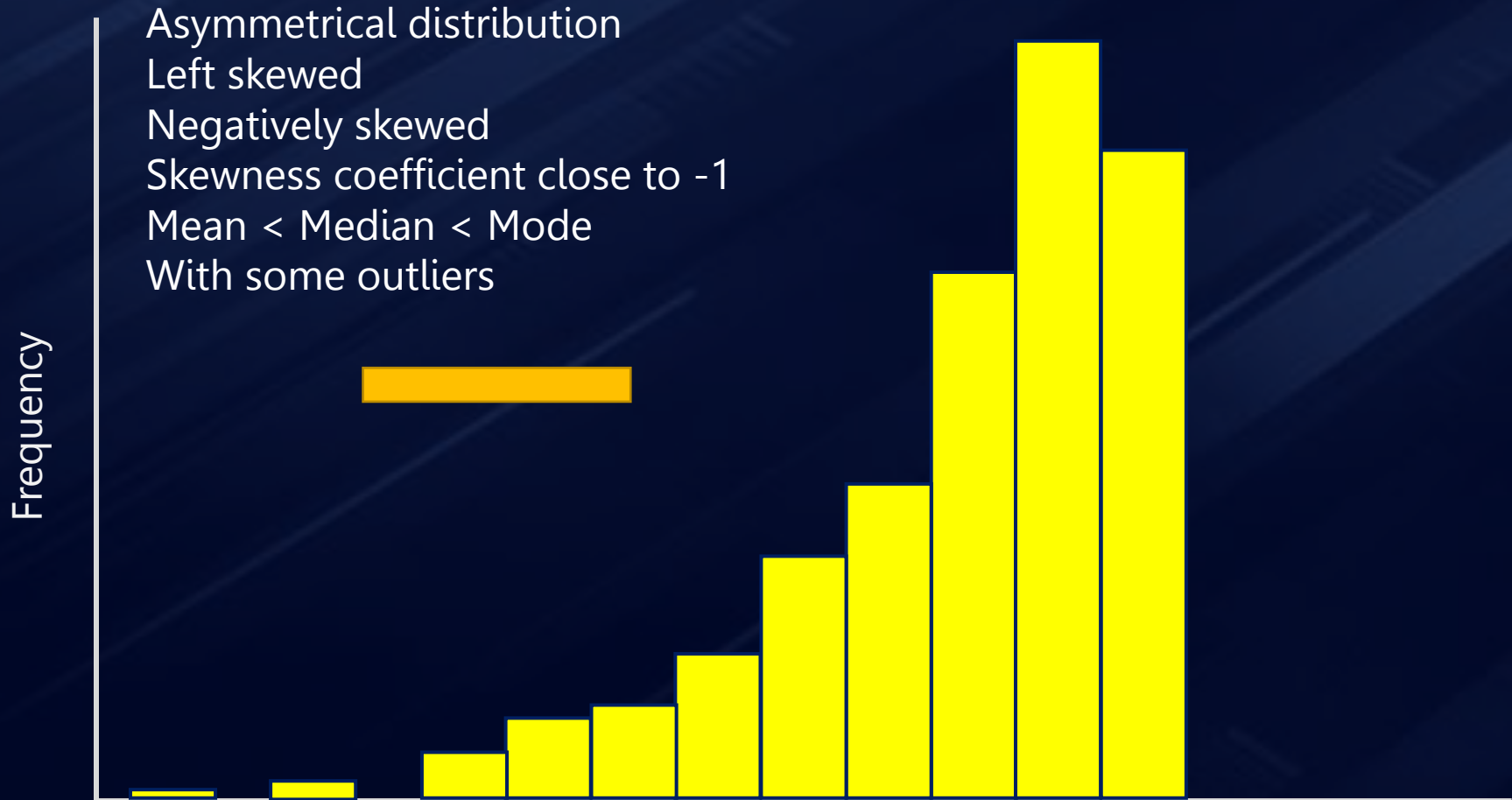
# Histogram

Frequency

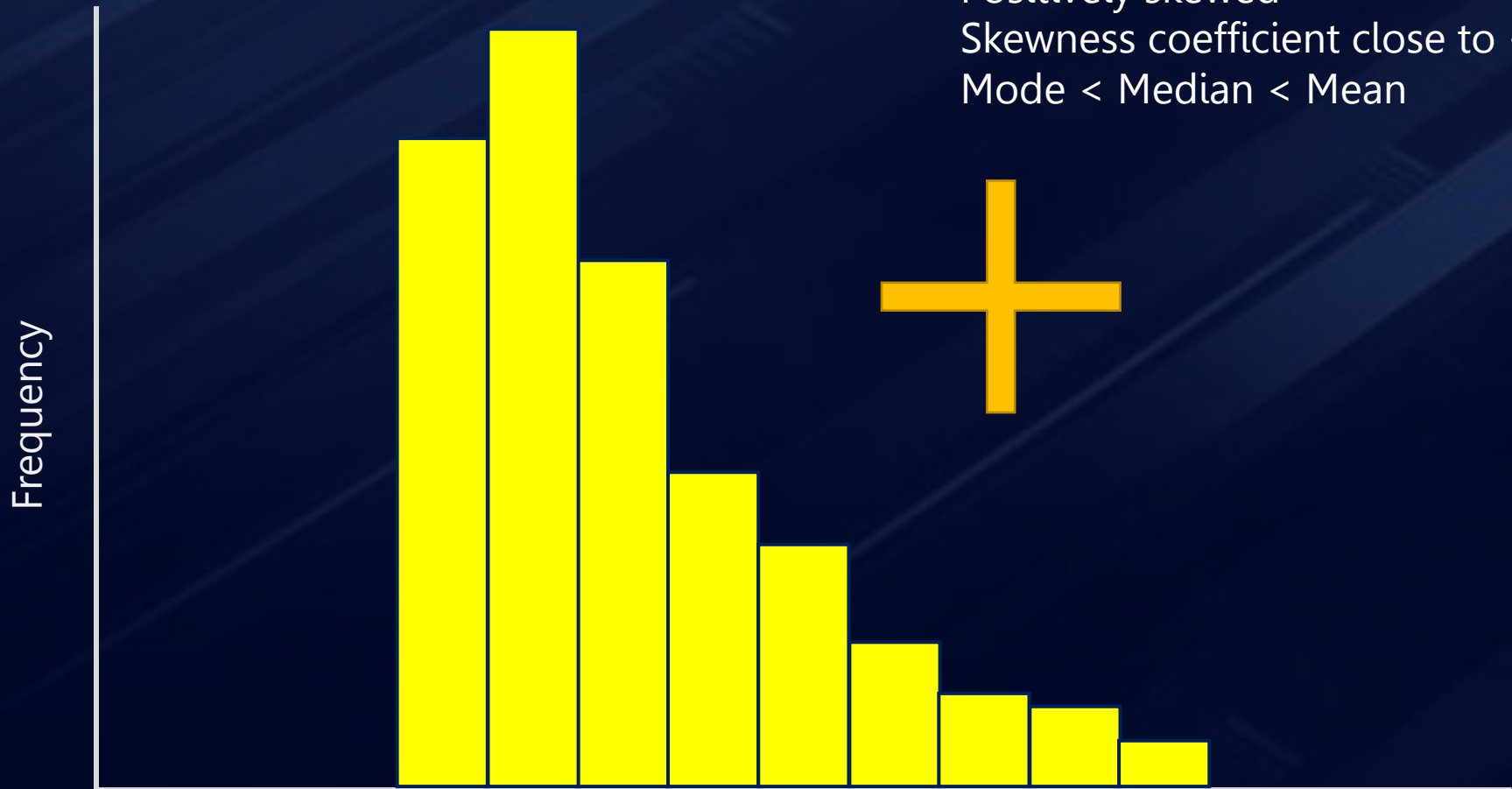
Symmetrical distribution  
Normal distribution  
Skewness near 0  
Mean = Median = Mode



# Histogram

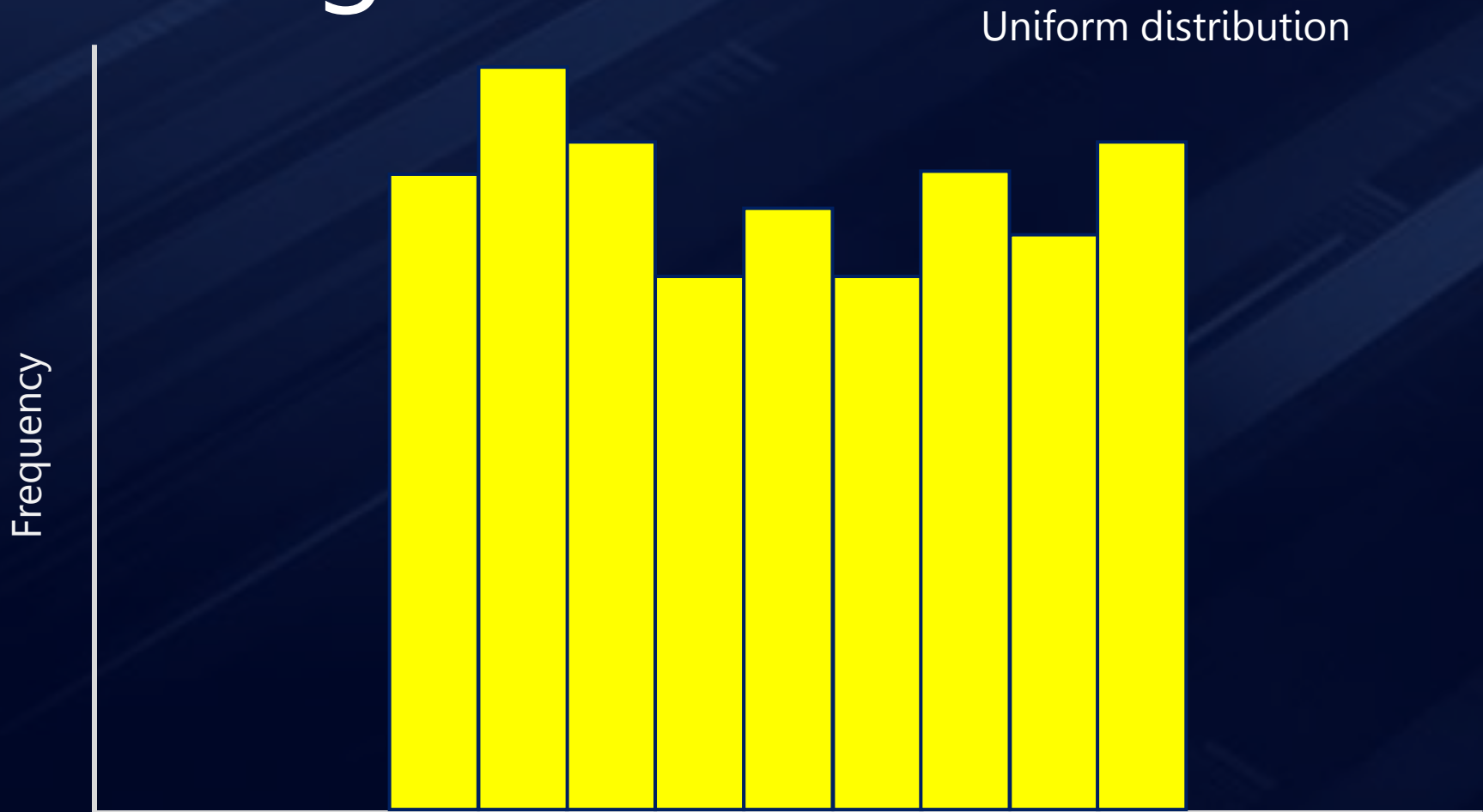


# Histogram



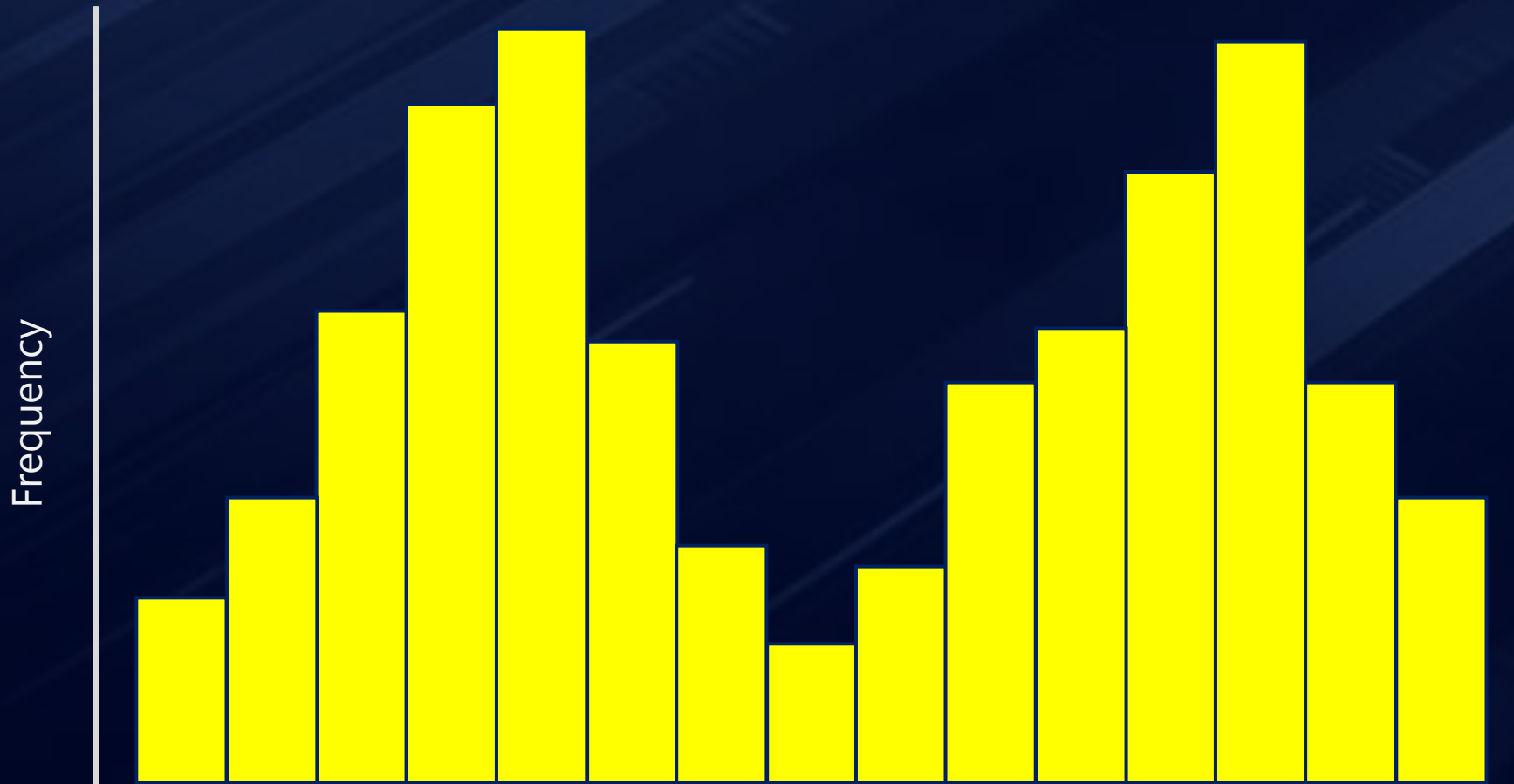
Asymmetrical distribution  
Right skewed  
Positively skewed  
Skewness coefficient close to +1  
Mode < Median < Mean

# Histogram



# Histogram

Bimodal distribution



# Frequency distribution

Age groups	n (%)
<15	0.2%
15-29	5%
30-39	7%
40-49	11%
50-59	16%
60-69	30%
70-79	35%
>80	14%

This table shows the age distribution of 2454 patients with acute pulmonary embolism. Source: modified from Goldhaber et al. (1999)

**What shape is the distribution?**

- a) Normal distribution
- b) Positively skewed distribution
- c) Negatively skewed distribution
- d) Uniform distribution
- e) Multimodal distribution



# Frequency distribution

- Measure of shape

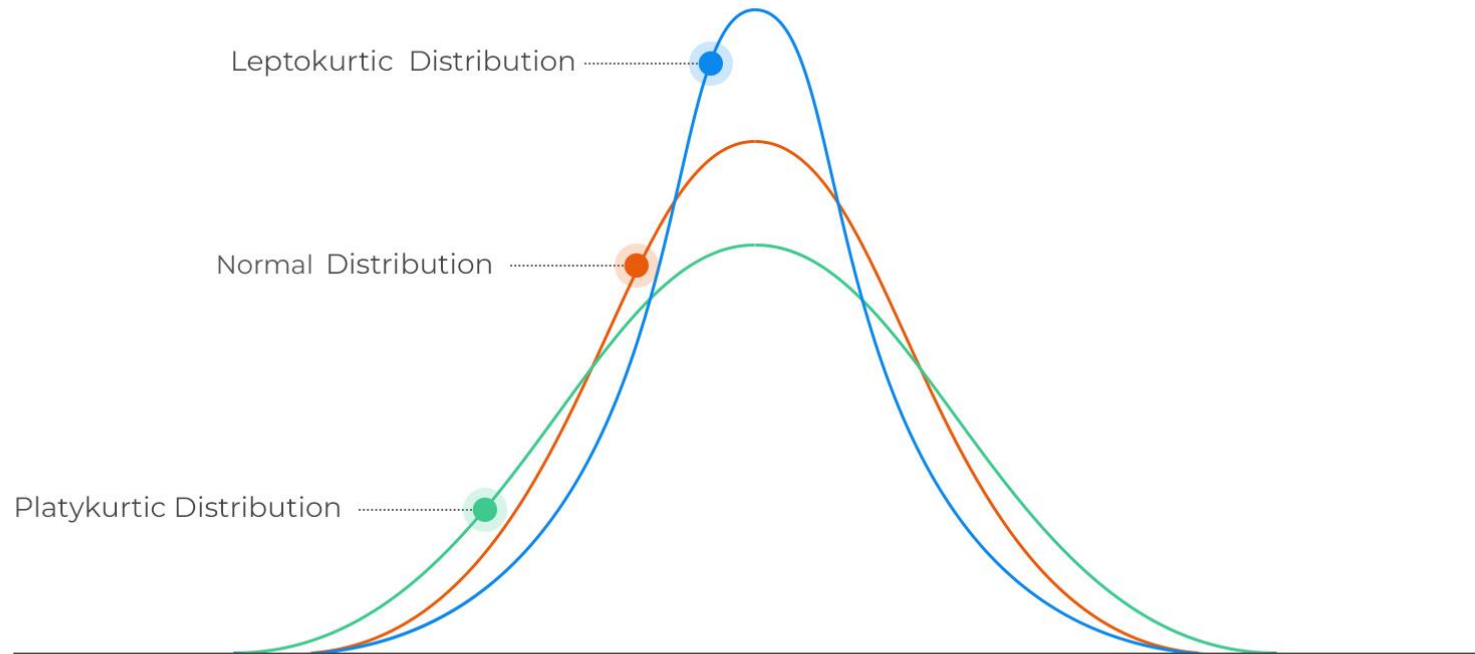
- Kurtosis

- Apply to symmetrical distribution
    - Whether the distribution has, to its left and right, short tails or long tails.
    - **High kurtosis value** = long tails (tails are longer and fatter, central peak is higher and more pointy)
    - **Low kurtosis value** = short tails (tails are shorter and thinner, central peak is lower and wider)

# Frequency distribution



## Kurtosis



<https://analystprep.com/cfa-level-1-exam/quantitative-methods/kurtosis-and-skewness-types-of-distributions/>

# Summarizing data

- Learning objectives
  - Measure of central tendency
    - Mode
    - Median
    - Mean
    - Minimum
    - Maximum
    - Percentile
    - Interquartile range

# Mode

- Mode, or modal values, is the value in the data with the **highest frequency**.
- Not actually a measure of central-ness.
- May be useful for nominal and ordinal data
- It is not useful for metric data types
- There may be more than one mode.
- Not very common in clinical research

# Median

- A measure of **central-ness** of the data.
- If we arrange the data in ascending order, the median is the **middle value**.
- Median is meaningless for nominal data.

- **Example I**

30 31 **32** 33 35

- **Example II**

30 31 **32 33** 35 50

# Median

- **Property of the median**
  - Not affected much by skewness of the distribution (good representative)
  - Not affected by outliers
  - Consider a very stable measure of centerness
- **Disadvantages**
  - Discards lots of information
  - Not easy to determine by hand

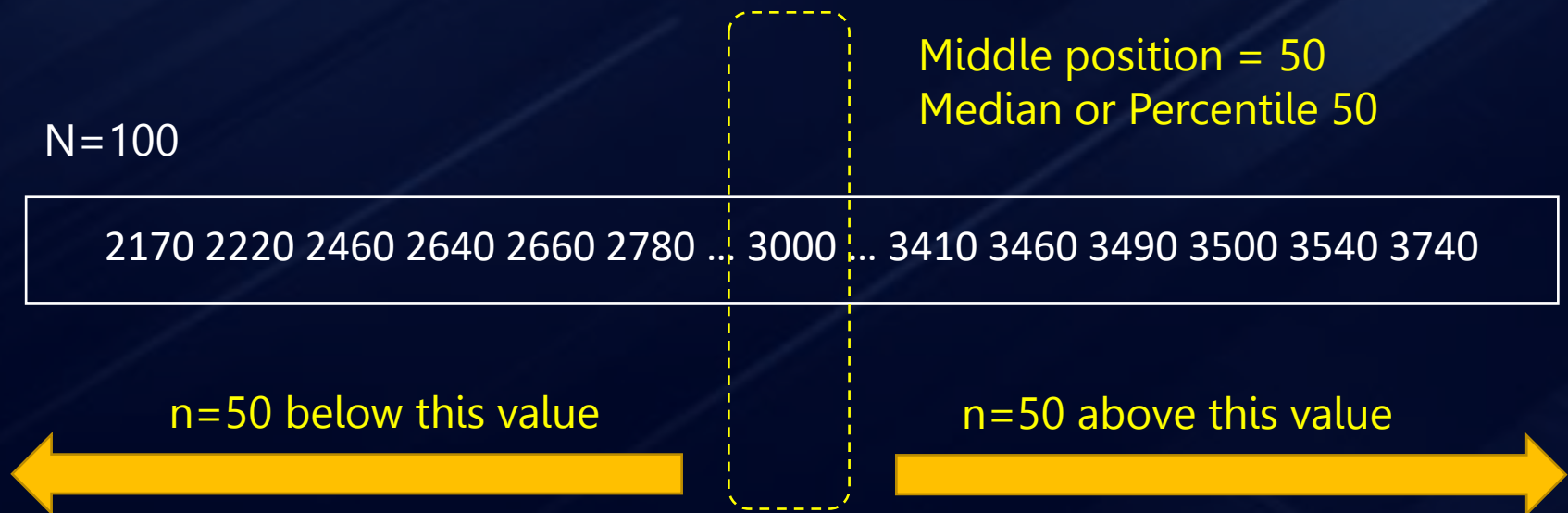


# Mean

- **Mean or arithmetic mean** or the average
- It uses all the information in the data to estimate
- It is affected by skewness and by outliers
- Mean might not be a good representative in some occasions.
- It should not be used with ordinal data (although very common in practice)
- **Example:** 30 31 32 33 32      Mean = 31.6
- **Example:** 30 31 32 33 32 70      Mean = 38.0

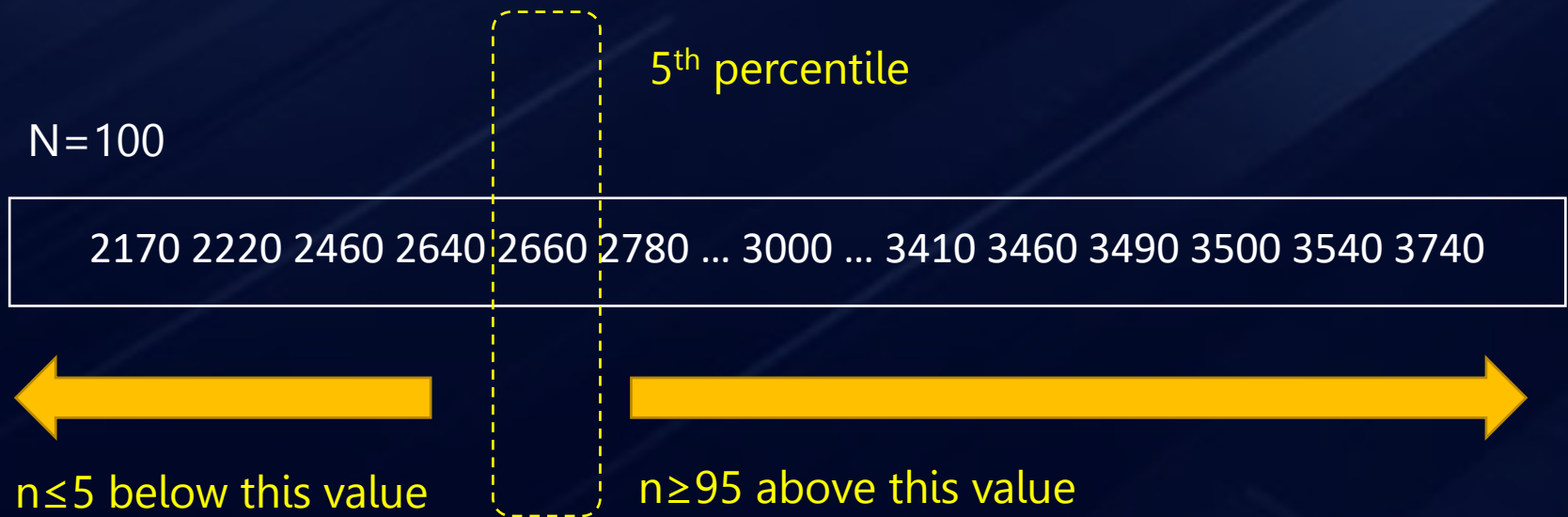
# Percentiles

- Are values which divide an ordered set of data into 100 equal-sized groups.



# Percentiles

- Are values which divide an ordered set of data into 100 equal-sized groups.



# What is appropriate?

- Depends on data type and distribution.

Type of data	Mode	Median	Mean
Nominal	Yes	No	No
Ordinal	Yes	Yes	Should not
Discrete	Yes	Yes if skewed	Yes if Normal
Continuous	No	Yes if skewed	Yes if Normal

# Summarizing data

- Learning objectives
  - Measure of dispersion (spread)
    - Range
    - Interquartile range
    - Standard deviation
    - Transformation of data

# Range

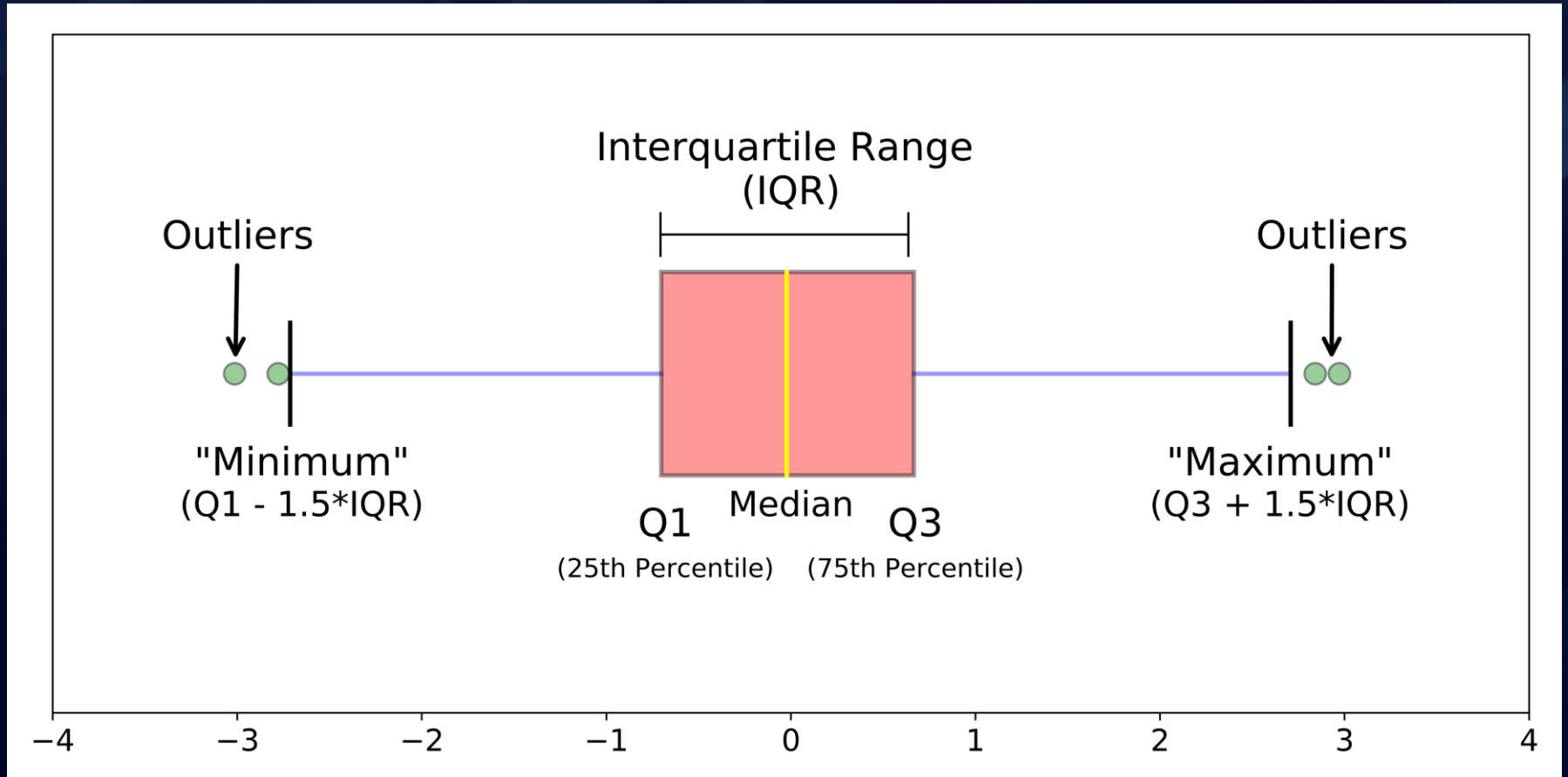
- Distance from the smallest value to the largest.
- Minimum to maximum
- Not affected by skewness
- Very sensitive to outliers in both ends.
- **Example:** 10 40 41 42 43 43 44 45 46 47 50 90 100
- Range (10 to 100)



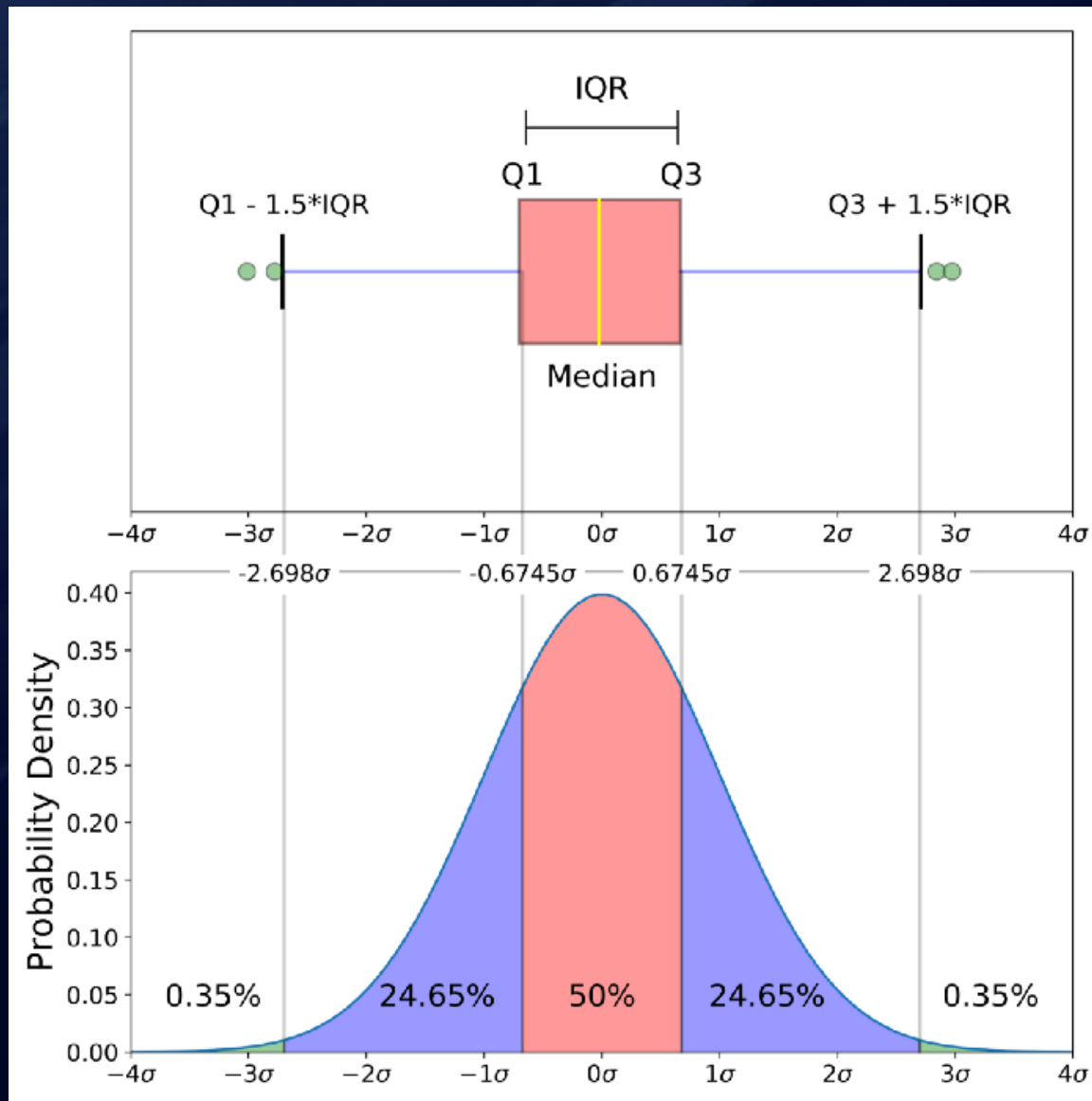
# Interquartile range (IQR)

- A solution to the problem of the sensitivity of the range to outliers in both ends.
- Slitted the data into 4 equal-sized quartiles
- 25%    25%    25%    25%
- Q1    Q2    Q3    Q4
- Then, chop a quarter of the values off both ends.
- IQR is written as Q1-Q3
- Not affected by outliers but can be affected by skewness. It does not use all the information in the data.

# Box and whisker plot



<https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

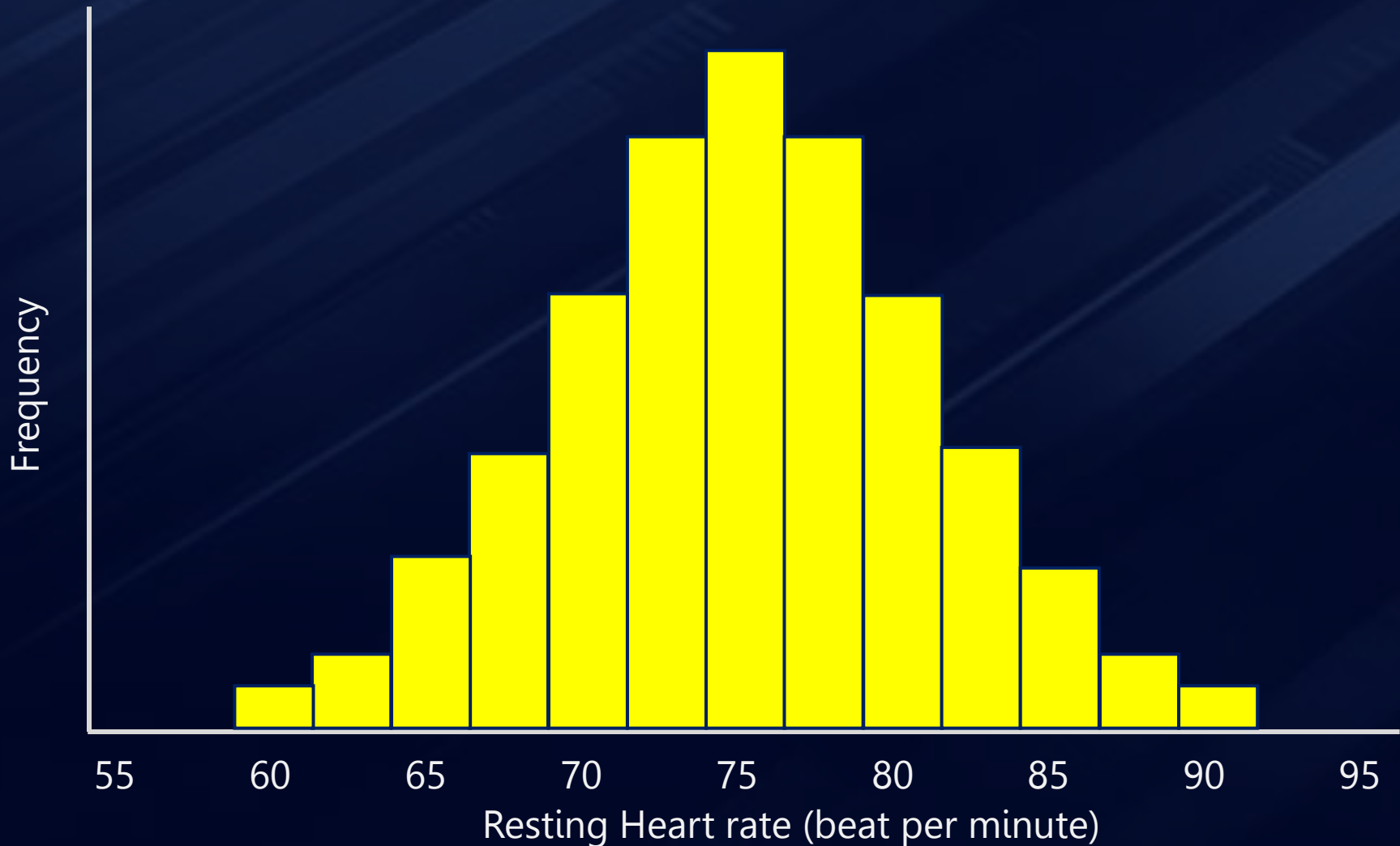


<https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

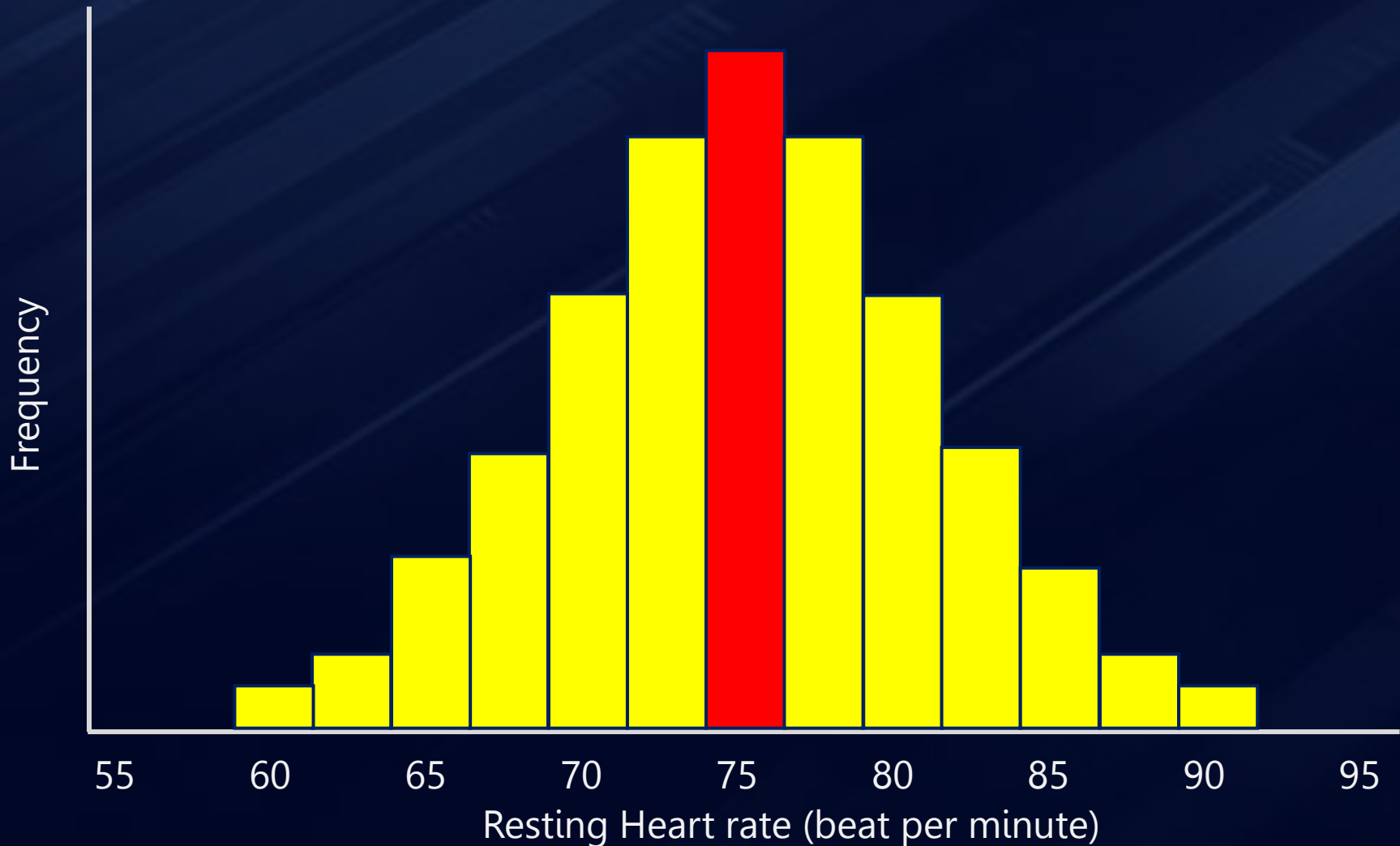
# Standard deviation

- To correct the limitation of IQR (not using all information in the data), the **standard deviation** is used instead.
- Can only be meaningfully used for **metric data**.
- Measure of the **average distance/deviation** of the data values from their collective mean.

# Standard deviation

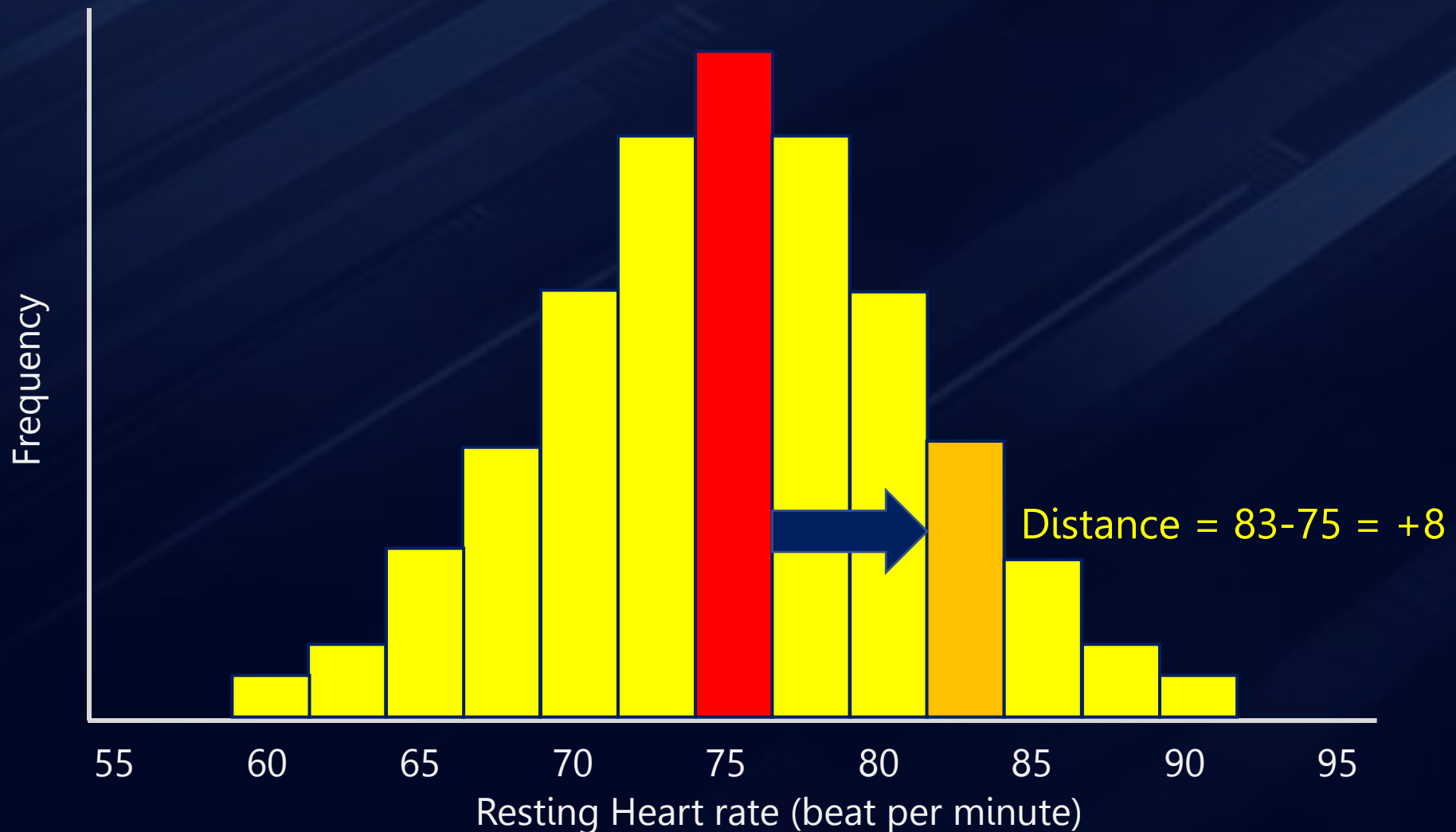


# Standard deviation

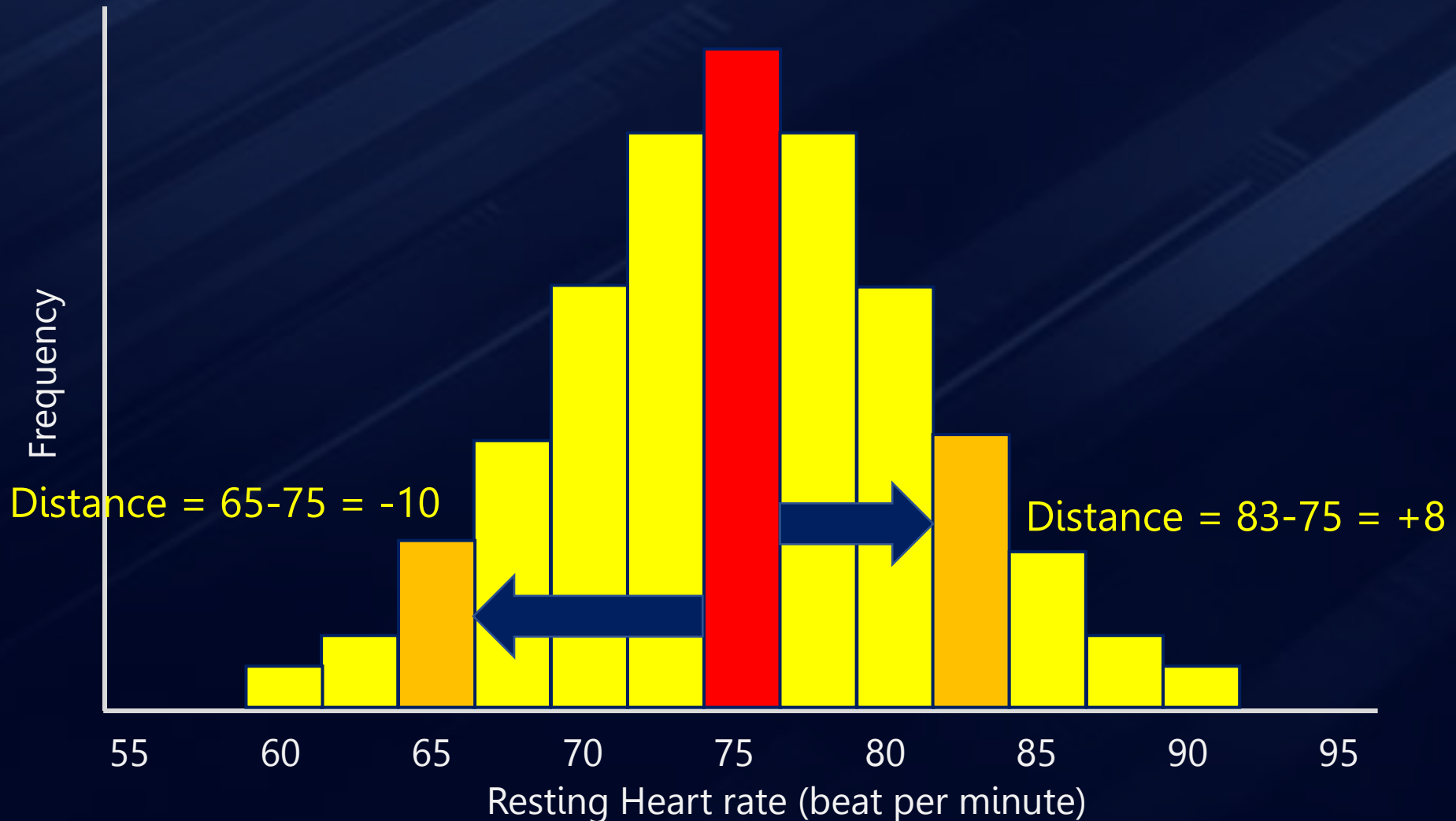




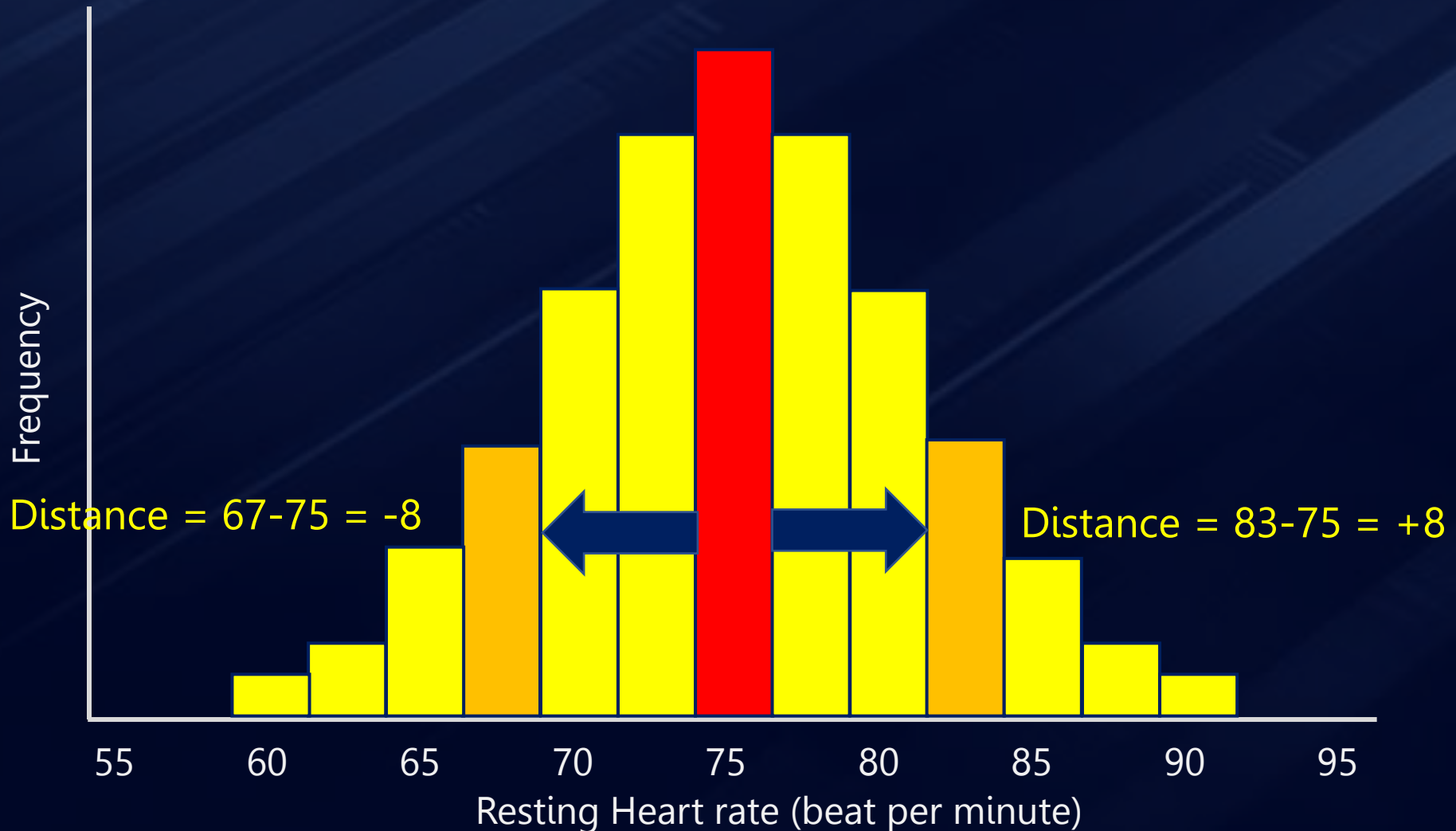
# Standard deviation



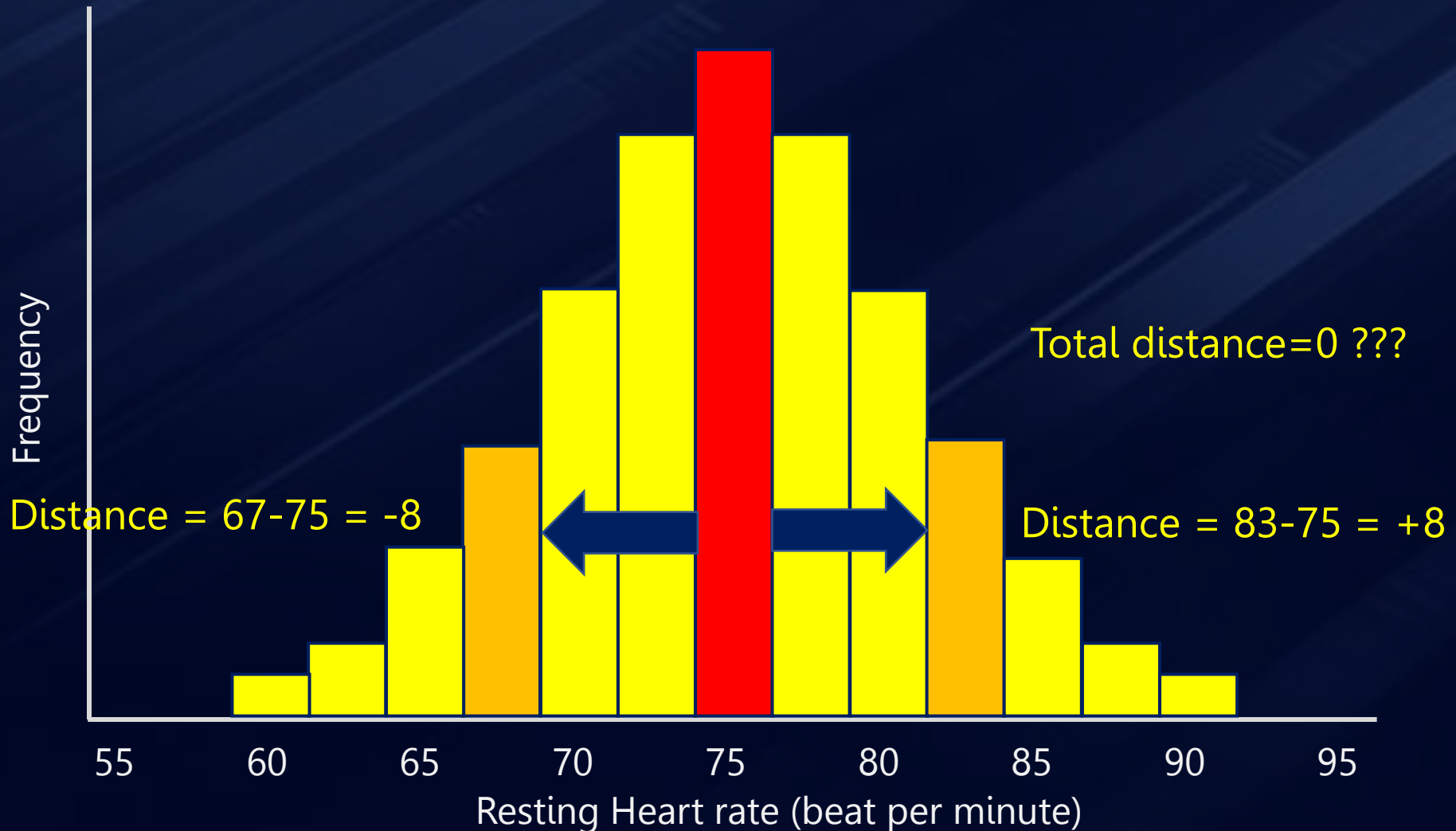
# Standard deviation



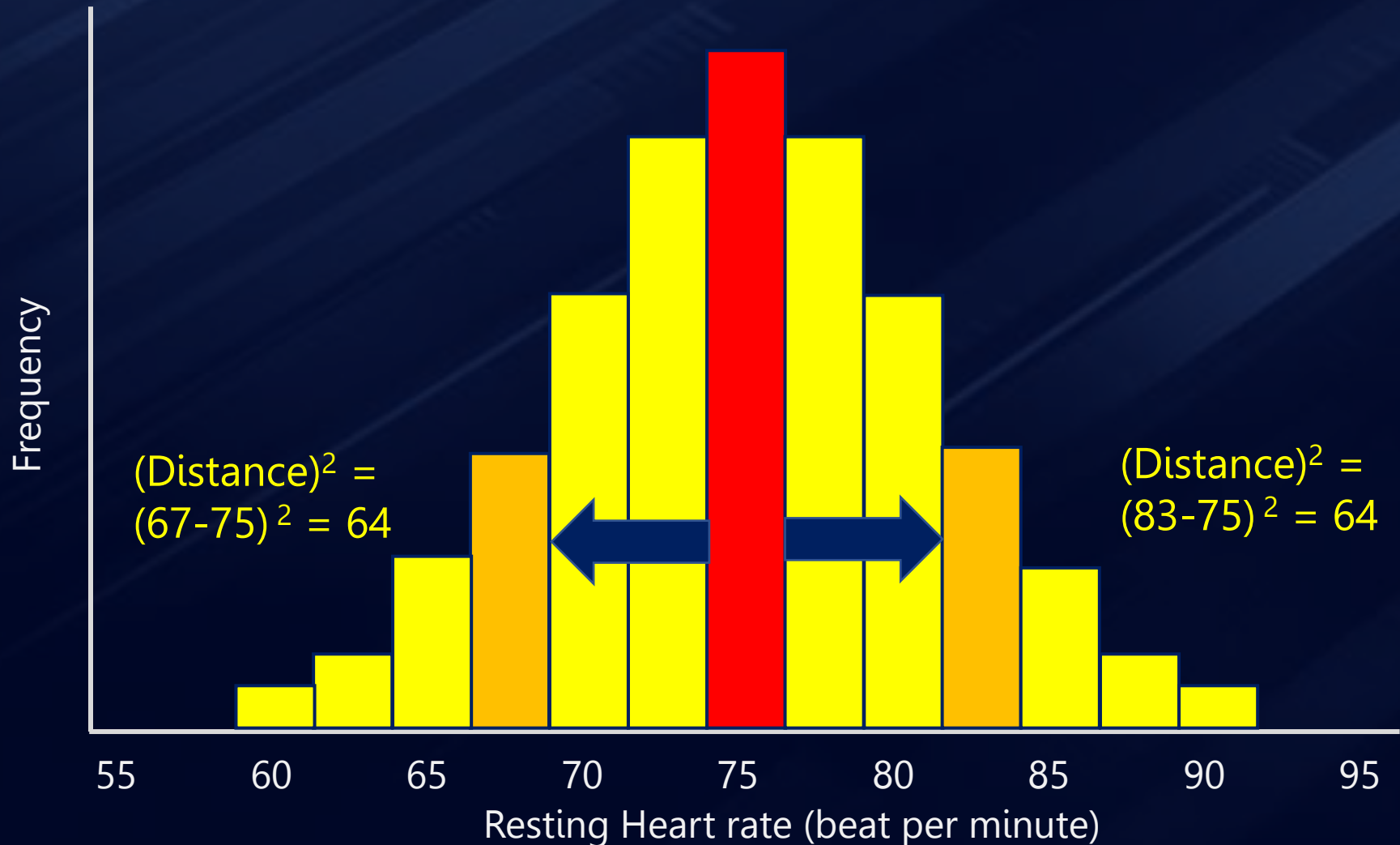
# Standard deviation



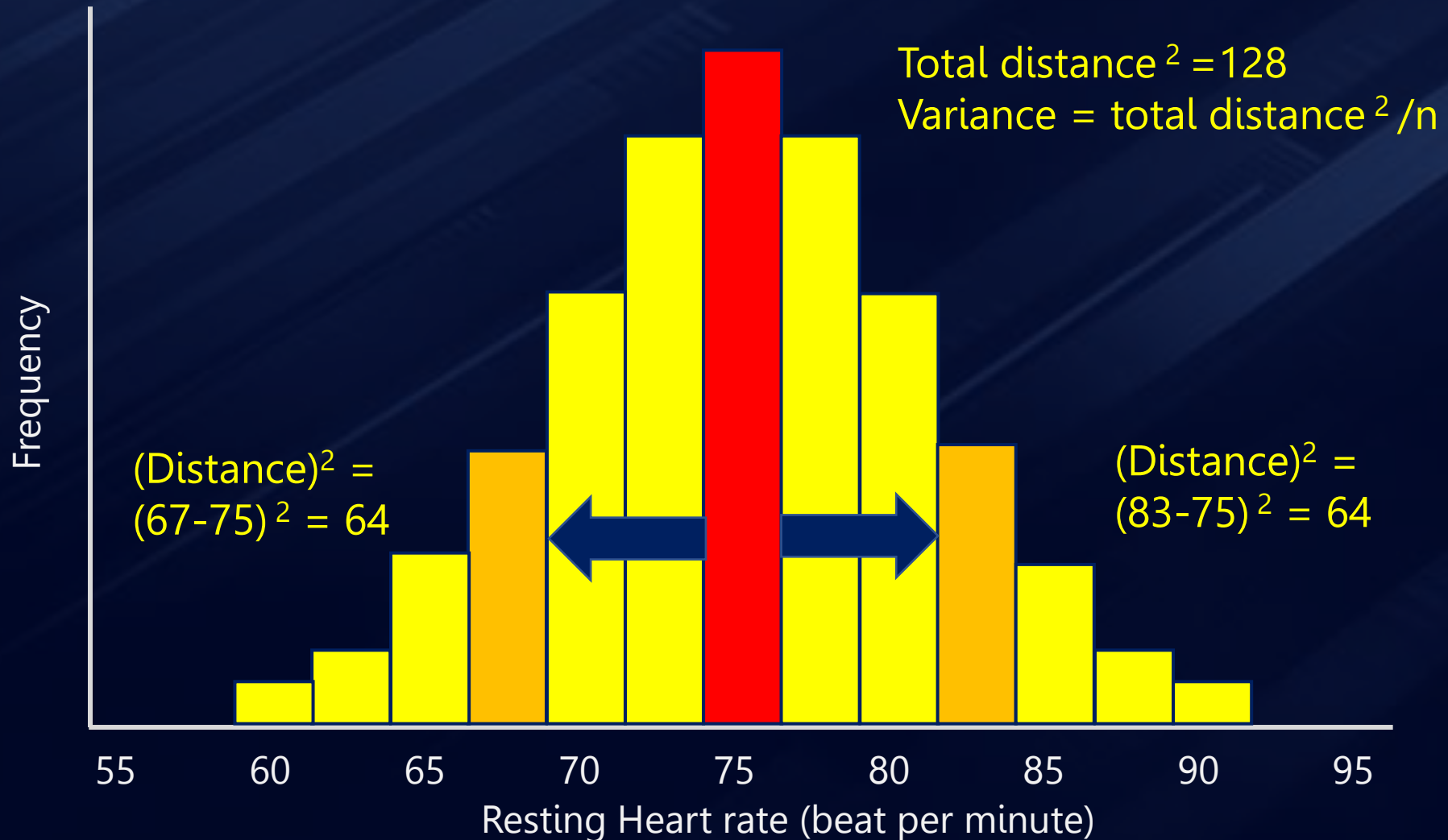
# Standard deviation



# Standard deviation



# Standard deviation





# Variance and SD



- $$\text{Variance} = \frac{(\text{Total distance from mean})^2}{n}$$

The unit is also squared; **beat per min**<sup>2</sup>

- $$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

The unit is back to **beat per min**

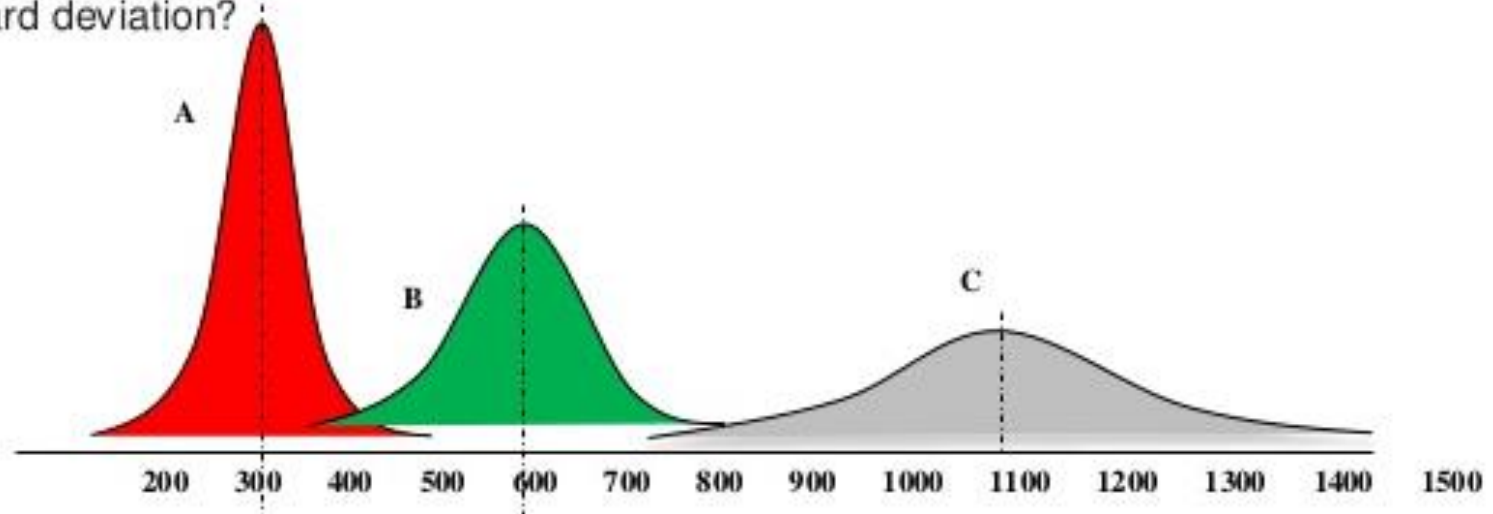
# Standard deviation

- Data spread  $\Rightarrow$  Distance from mean  $\Rightarrow$  SD
- If the **data are widely spread**, the average distance of the values from their mean will be large, and thus the **standard deviation**. 
- If the **data are narrowly spread**, the average distance of the values from their mean will be small, and thus the **standard deviation**. 

# Standard deviation

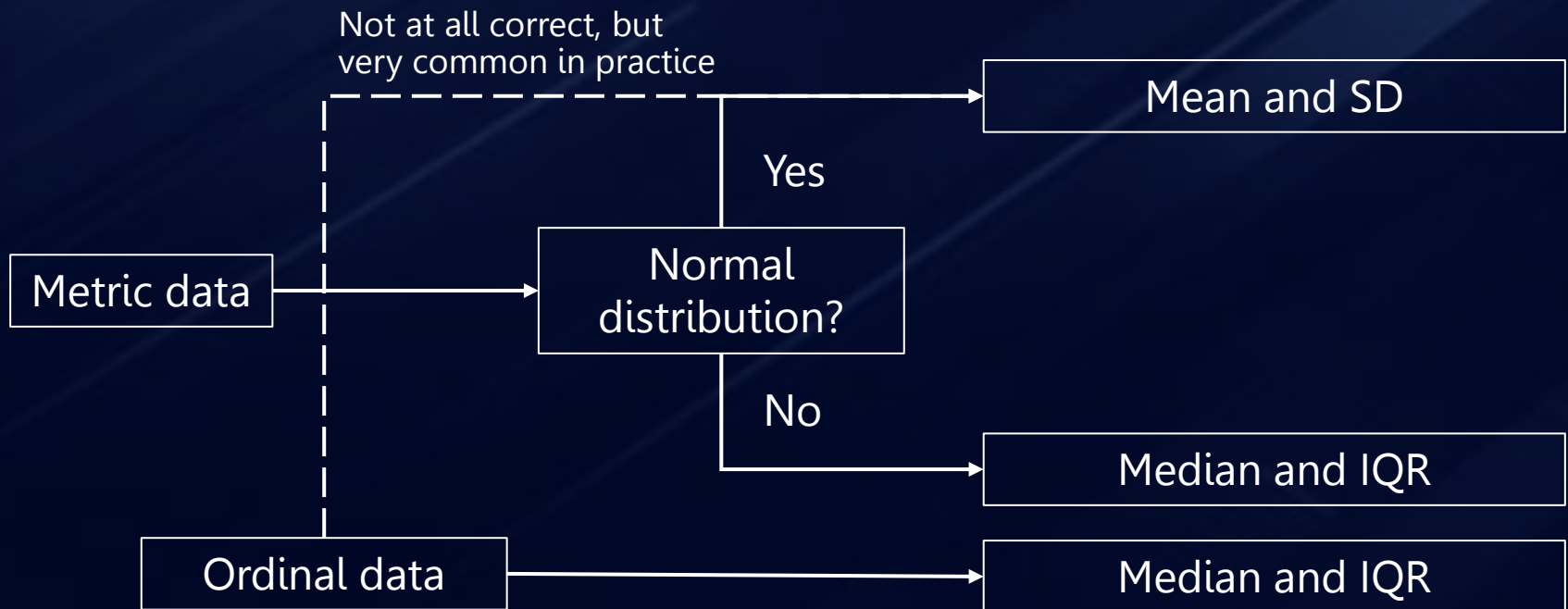
**Working Problem 6.6:** Given the three normal distributions A, B, and C below:

- Which normal curve has the greatest mean and which has the lowest mean?
- Which normal curve has the greatest standard deviation and which has the lowest standard deviation?



# Standard deviation

- **Consideration**
  - Only is meaningful if the data is **normally distributed**, hence the use of mean.



# Standard deviation

- Interpretation?

**Table 1.** Baseline clinical characteristics and potential predictors for all-cause mortality in Thai patients with fragility fracture of hip.

Characteristics	Dead <i>n</i> , (%) ( <i>n</i> = 108)	Alive <i>n</i> , (%) ( <i>n</i> = 667)	<i>p</i> -Value
<b>General Factors</b>			
Gender, <i>n</i> (%)			
Male	36 (33.33)	178 (26.69)	0.164
Female	72 (66.67)	489 (73.31)	
Age (years), Mean $\pm$ SD <sup>c</sup>	81.63 $\pm$ 8.52	78.68 $\pm$ 9.65	0.003
Age at admission $\geq$ 85 years, <i>n</i> (%)	46 (42.59)	188 (28.19)	0.003
BMI <sup>a</sup> at admission (kg/m <sup>2</sup> ), Mean $\pm$ SD <sup>c</sup> ( <i>n</i> = 769)	19.82 $\pm$ 3.13	21.18 $\pm$ 4.05	<0.001
BMI <sup>a</sup> at admission $\geq$ 25 kg/m <sup>2</sup> , <i>n</i> (%) ( <i>n</i> = 769)	8 (7.48)	96 (14.50)	0.048
Pre-fracture walking ability by oneself, <i>n</i> (%) ( <i>n</i> = 606)	72 (91.14)	504 (95.64)	0.095
Living with family, <i>n</i> (%) ( <i>n</i> = 774)	108 (100)	663 (99.55)	1.000

Atthakomol et al. (2020)



# Standard deviation

- Interpretation?

**Table 1.** Baseline clinical characteristics and potential predictors for all-cause mortality in Thai patients with fragility fracture of the hip

Characteristic	Number of patients (%)		P-Value
Gender			
Male	36 (3.7)	178 (26.69)	0.164
Female	72 (6.3)	489 (73.31)	
Age (years), Mean $\pm$ SD <sup>c</sup>	81.63 $\pm$ 8.52	78.68 $\pm$ 9.65	0.003
Age at admission $\geq$ 85 years, <i>n</i> (%)	46 (42.59)	188 (28.19)	0.003
BMI <sup>a</sup> at admission (kg/m <sup>2</sup> ), Mean $\pm$ SD <sup>c</sup> ( <i>n</i> = 769)	19.82 $\pm$ 3.13	21.18 $\pm$ 4.05	<0.001
BMI <sup>a</sup> at admission $\geq$ 25 kg/m <sup>2</sup> , <i>n</i> (%) ( <i>n</i> = 769)	8 (7.48)	96 (14.50)	0.048
Pre-fracture walking ability by oneself, <i>n</i> (%) ( <i>n</i> = 606)	72 (91.14)	504 (95.64)	0.095
Living with family, <i>n</i> (%) ( <i>n</i> = 774)	108 (100)	663 (99.55)	1.000

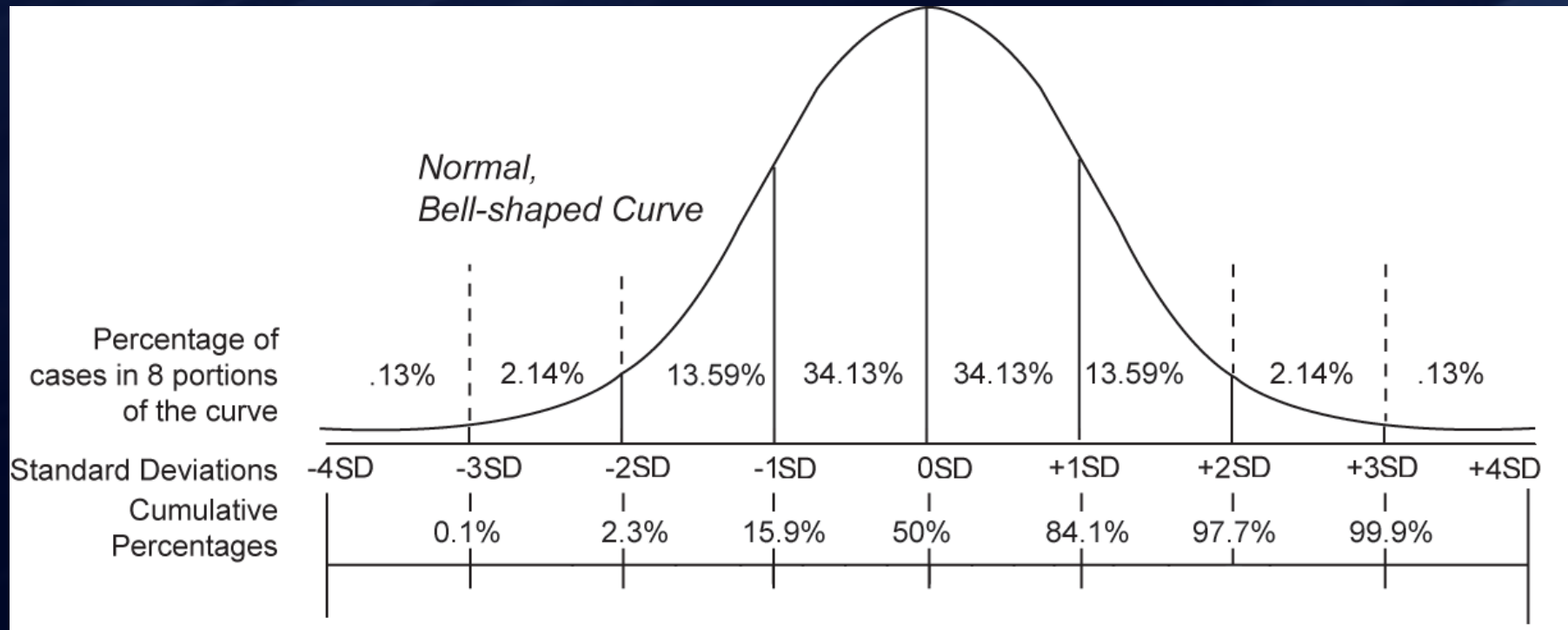
The average distance/deviation of the age values is about 8.5 years from the sample mean of 81.6 years

Atthakomol et al. (2020)



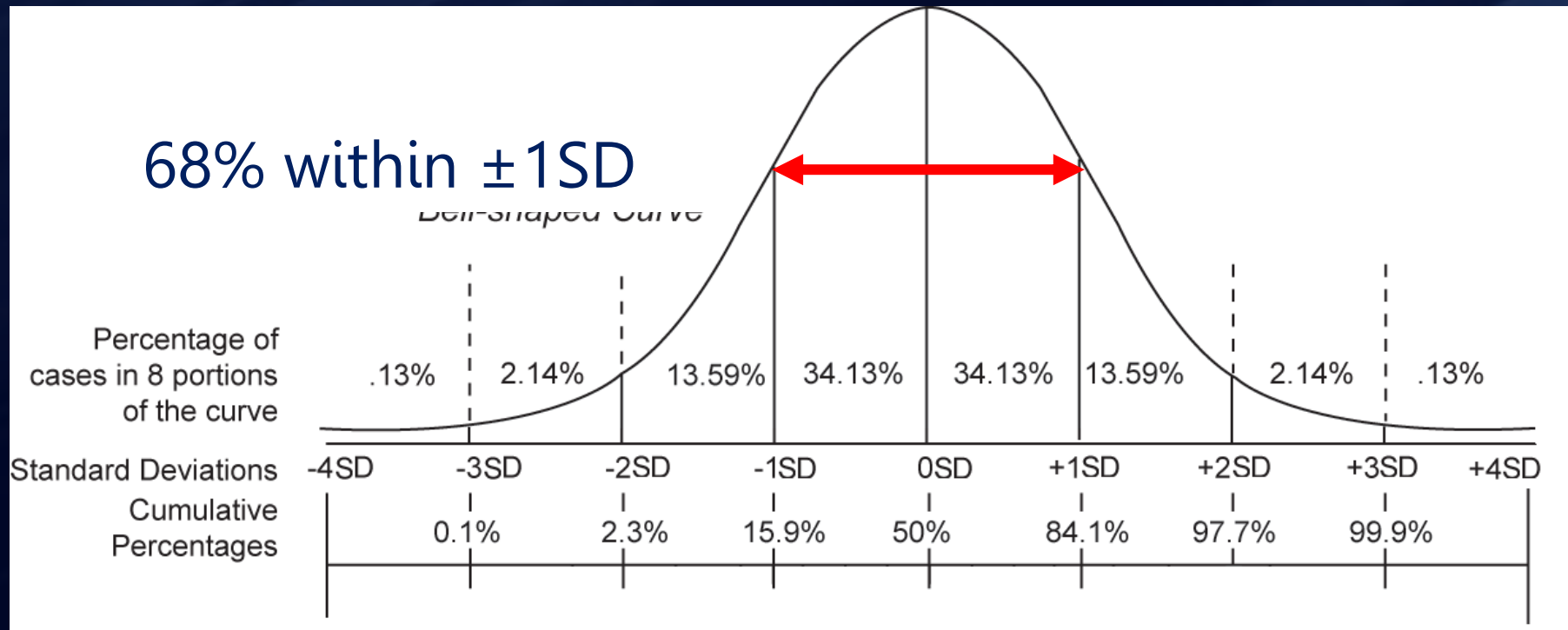
# Standard deviation

- Area properties of the Normal distribution



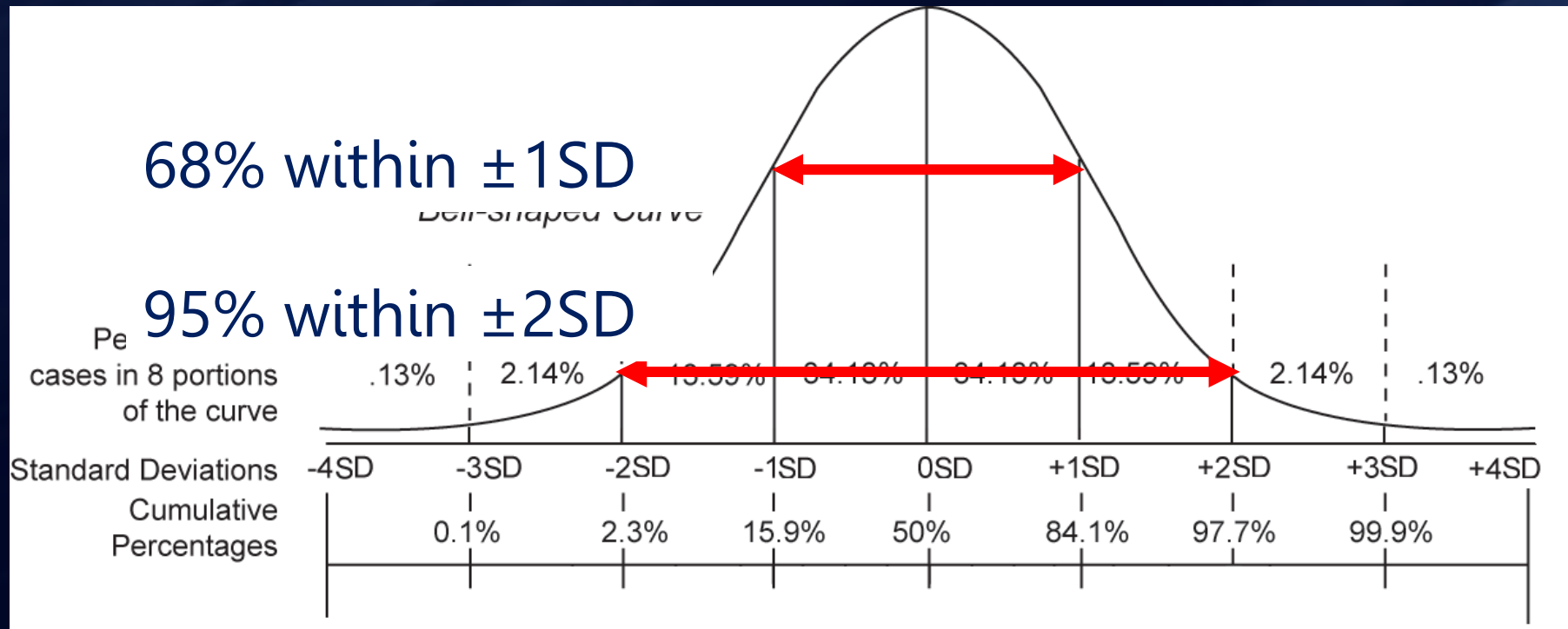
# Standard deviation

- Area properties of the Normal distribution



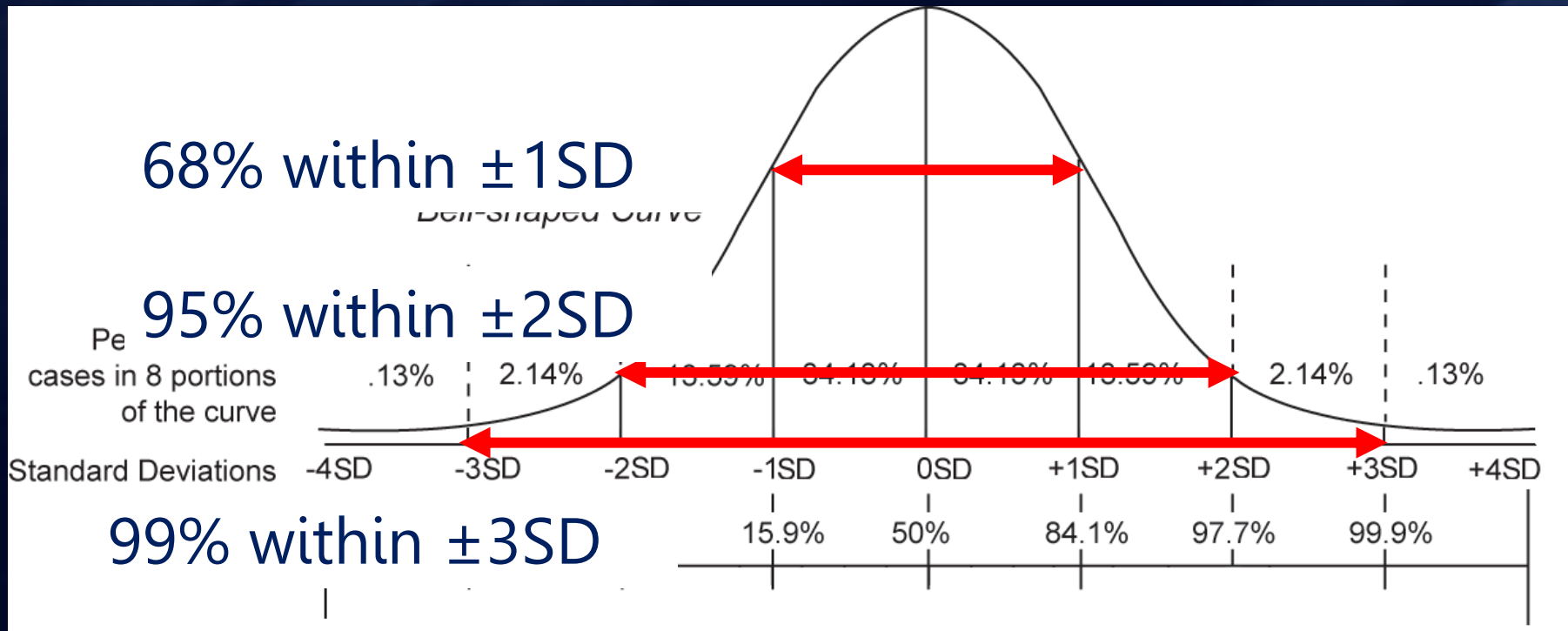
# Standard deviation

- Area properties of the Normal distribution



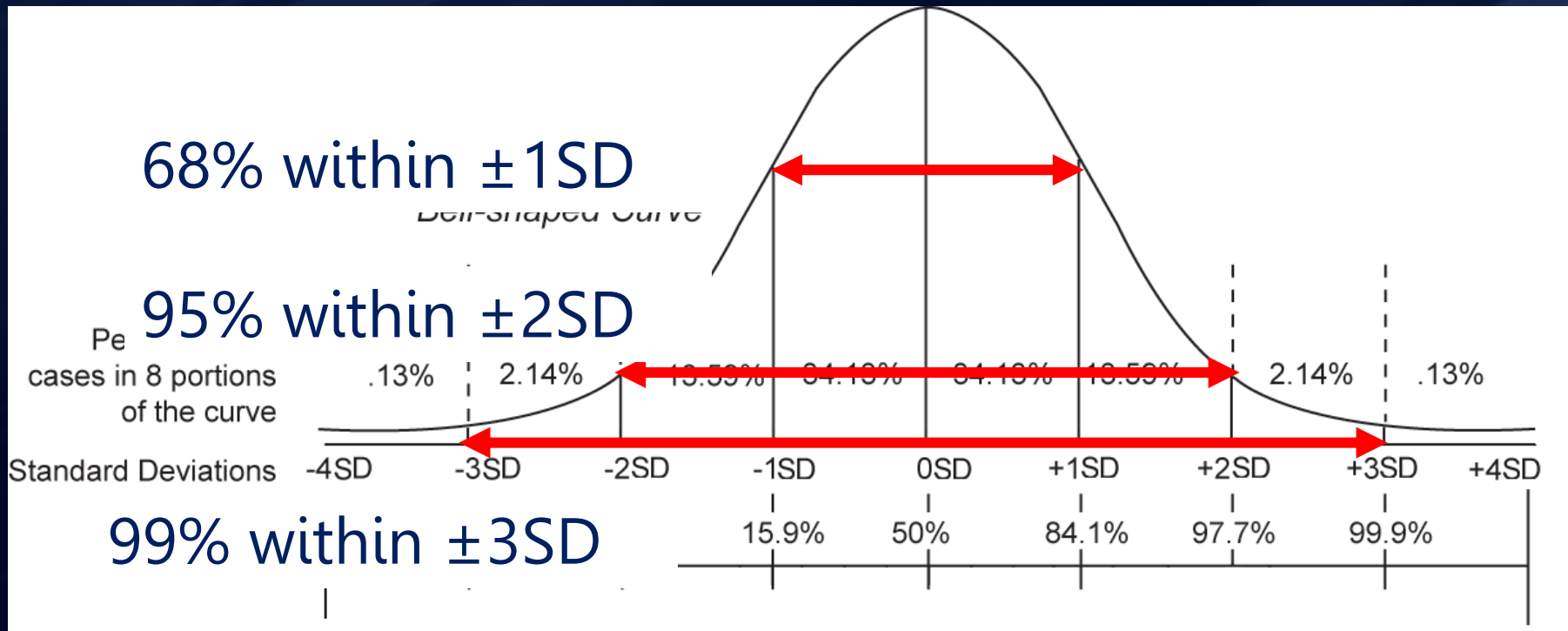
# Standard deviation

- Area properties of the Normal distribution



# Standard deviation

- Area properties of the Normal distribution



Assumption: The data distribution must be Normal!!

# Standard deviation

- Interpretation?

68% of the age values lie between 73.11 to 90.15 years  
95% of the age values lie between 64.6 To 98.67 years

**Table 1.** Baseline clinical characteristics of patients with fragility fracture of the hip

Characteristic	Patients (n = 769)	Controls (n = 769)	P-Value
Gender			
Male	36 (3.7)	178 (26.69)	0.164
Female	72 (6.37)	489 (73.31)	
Age (years), Mean $\pm$ SD <sup>c</sup>	81.63 $\pm$ 8.52	78.68 $\pm$ 9.65	0.003
Age at admission $\geq$ 85 years, n (%)	46 (42.59)	188 (28.19)	0.003
BMI <sup>a</sup> at admission (kg/m <sup>2</sup> ), Mean $\pm$ SD <sup>c</sup> (n = 769)	19.82 $\pm$ 3.13	21.18 $\pm$ 4.05	<0.001
BMI <sup>a</sup> at admission $\geq$ 25 kg/m <sup>2</sup> , n (%) (n = 769)	8 (7.48)	96 (14.50)	0.048
Pre-fracture walking ability by oneself, n (%) (n = 606)	72 (91.14)	504 (95.64)	0.095
Living with family, n (%) (n = 774)	108 (100)	663 (99.55)	1.000

Atthakomol et al. (2020)



# Normal distribution

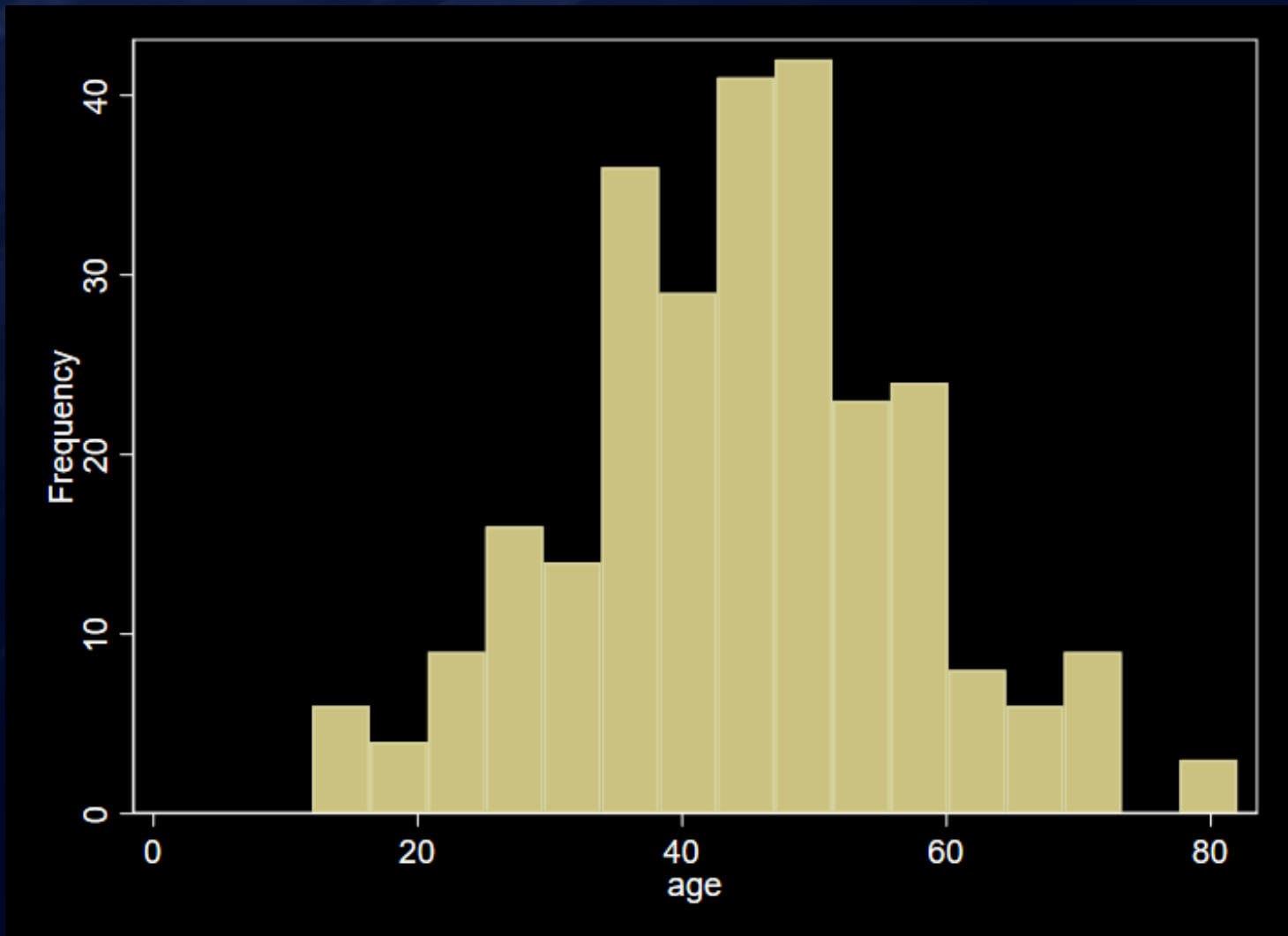
- Symmetrical bell-shaped distribution
- Special place in the heart of statisticians!
- What statisticians really want to know is whether or not a distribution is *Normal* or *Normal-ish*!

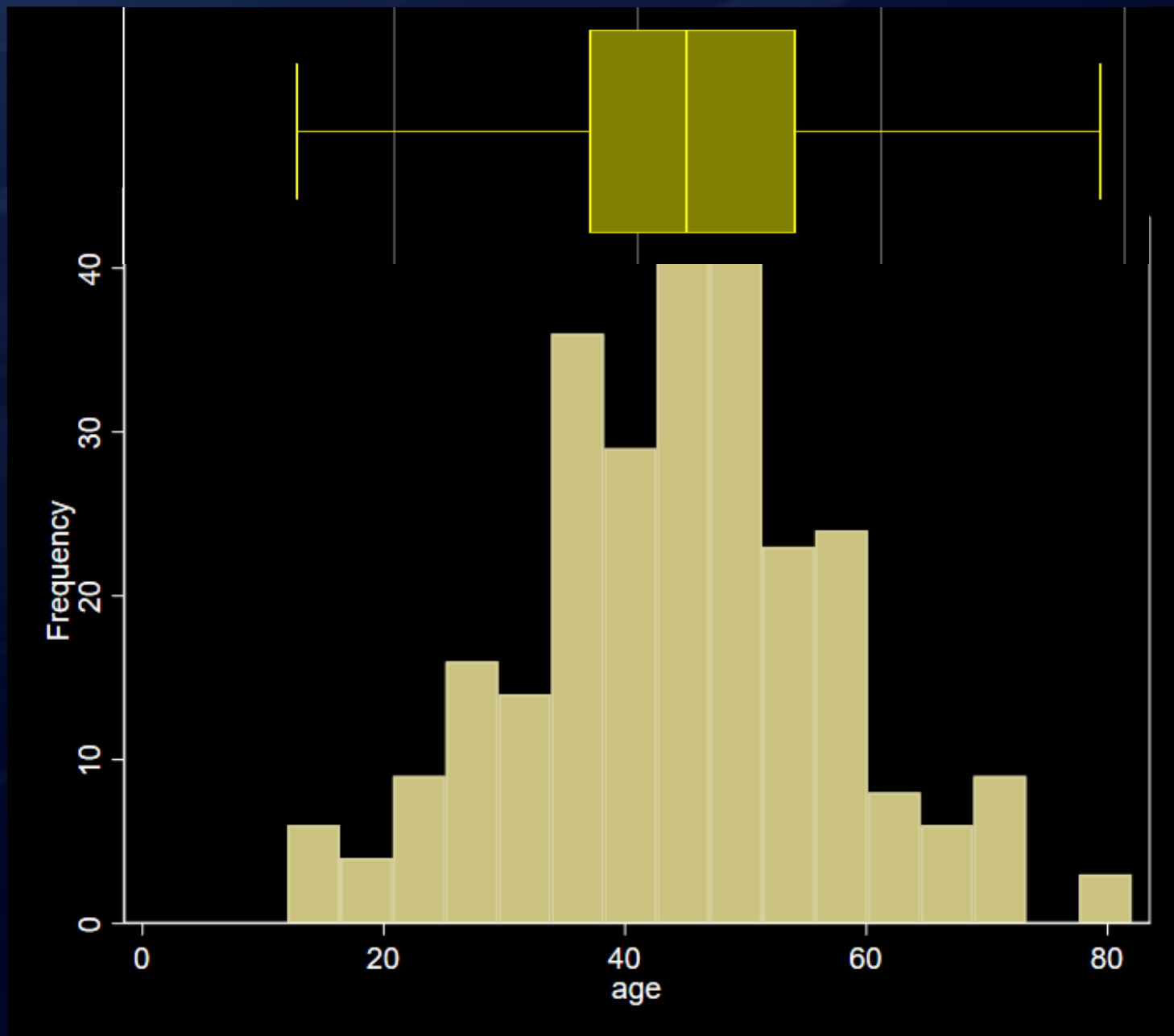


# Normal distribution

- Three key characteristics!
  - Visualization
    - Histogram
    - Box plot
  - Mean and median approximation
  - Large standard deviation
  - ~~Statistical tests for normality~~

# Normal distribution





# Normal distribution

```
. sum age,detail
```

age				
Percentiles		Smallest		
1%	15	12		
5%	22	14		
10%	27	15	Obs	270
25%	36	15	Sum of Wgt.	270
50%	44		Mean	44.5037
		Largest	Std. Dev.	12.99206
75%	53	73	Variance	168.7937
90%	60	78	Skewness	.0584846
95%	67	78	Kurtosis	3.078091
99%	78	82		

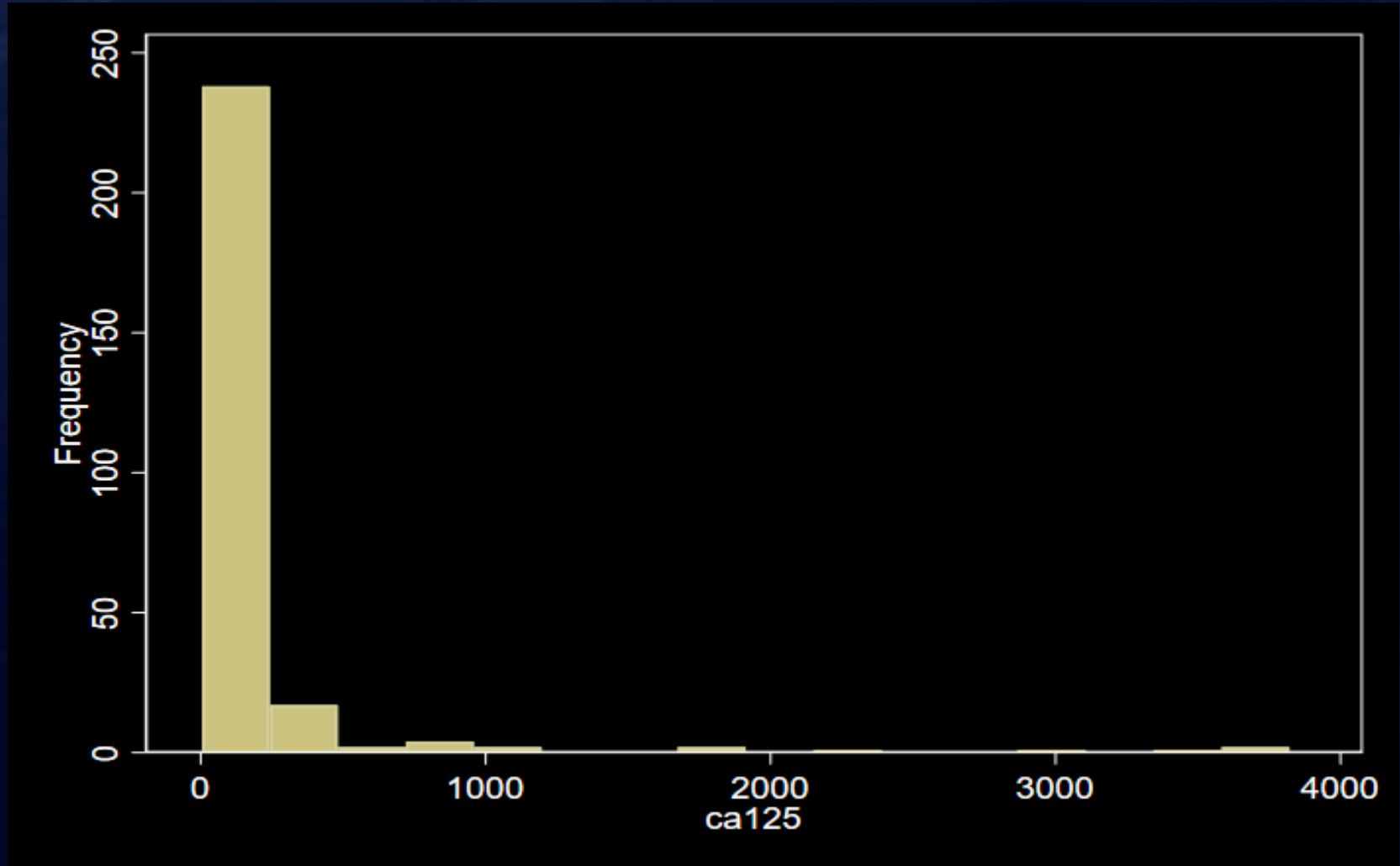
# Normal distribution ?????

```
. sum ca125,detail
```

ca125

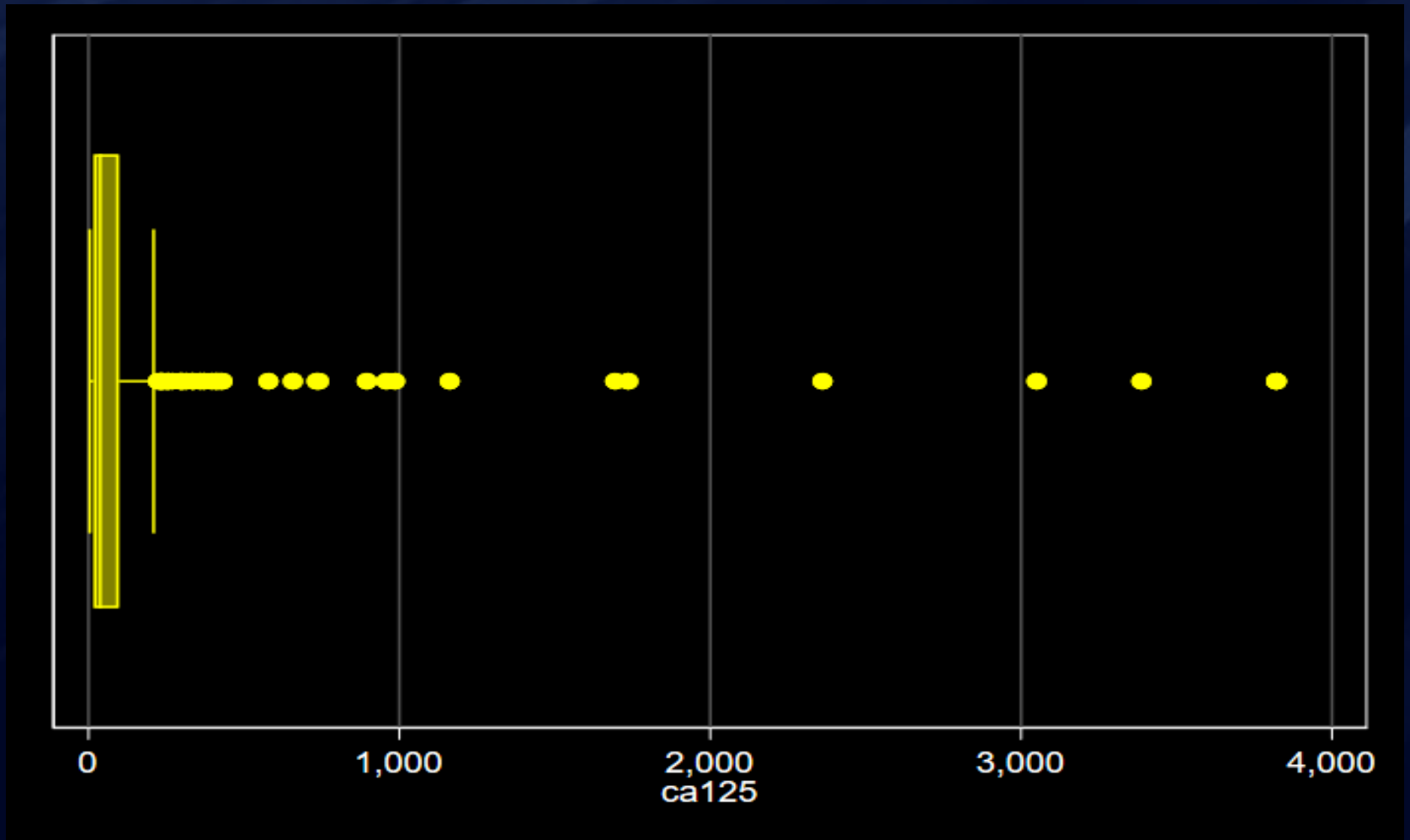
Percentiles		Smallest		
1%	5.96	3.97		
5%	9.12	5.85		
10%	10.44	5.96	Obs	270
25%	16.3	6.1	Sum of Wgt.	270
50%	37.54		Mean	165.2111
		Largest	Std. Dev.	484.8014
75%	97.93	3050		
90%	297.25	3387	Variance	235032.4
95%	657.2	3821	Skewness	5.746399
99%	3387	3821	Kurtosis	38.6065

# Positively skewed distribution

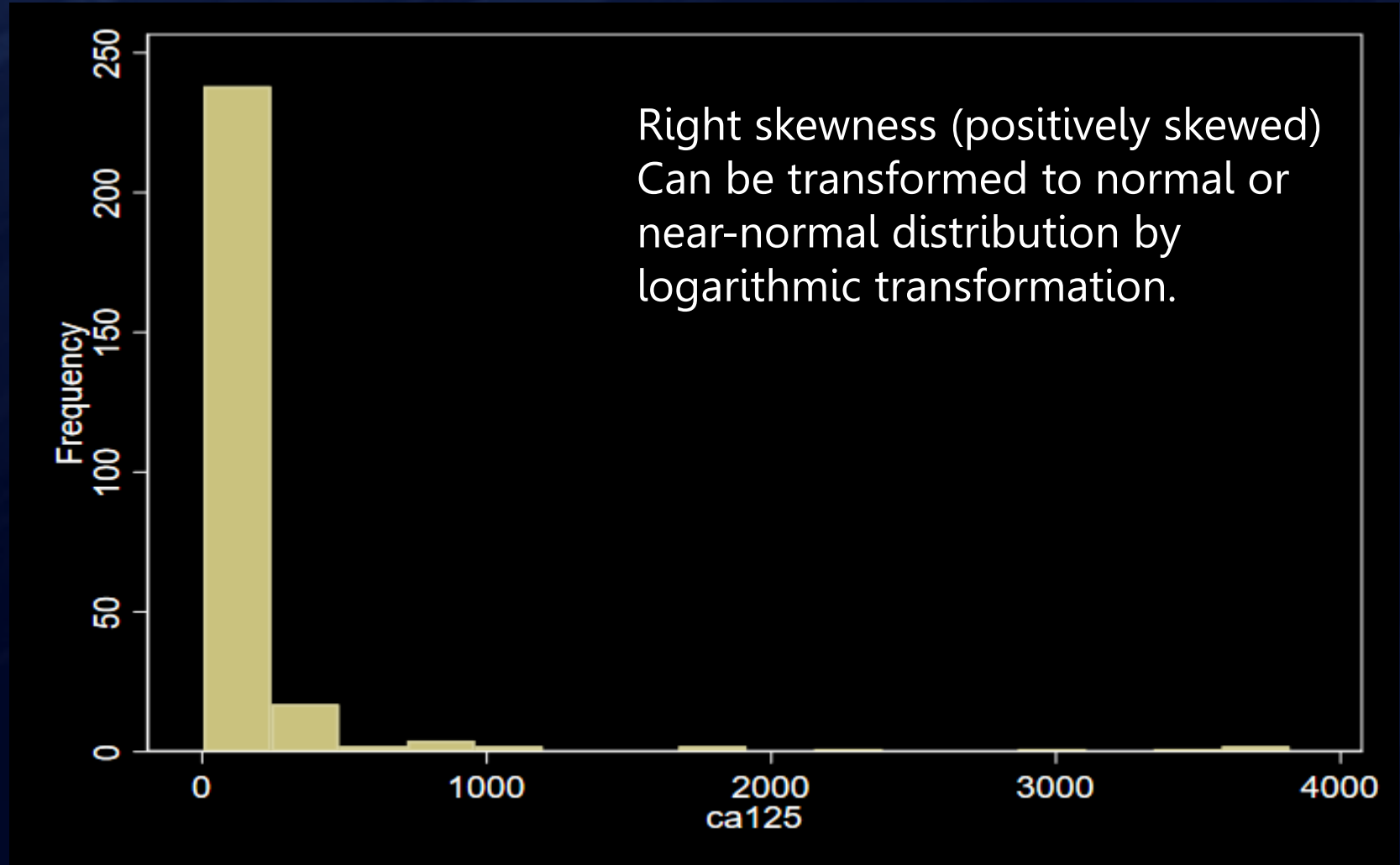




# Positively skewed distribution



# Transformation of data



# Transformation of data

